FITTING GRAPHS AND TREES WITH MULTIDIMENSIONAL SCALING METHODS

Willem J. Heiser

Department of Data Theory
Leiden University

RR-96-06

Fitting Graphs and Trees with Multidimensional Scaling Methods

Willem J. Heiser

Department of Data Theory, Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands

Summary: The symmetric difference between sets of qualitative elements (called features) forms the basis of a distance model that can be used as a general framework for fitting a particular class of graphs, which includes additive trees, hierarchical trees and circumplex structures. It is shown how to parametrize this fitting problem in terms of a lattice of subsets, and how inclusion relations between feature sets lead to additivity of distance along paths in a graph. An algorithm based on alternating least squares and on the recent method of cluster differences scaling is described, and illustrated for the general case.

1. Introduction: Fitting distances or coordinates

Graphs and trees are increasingly considered to be attractive discrete structures for modelling general similarity or dissimilarity (or: proximity) data in the social and behavioral sciences (Arabie and Hubert, 1992; Klauer, 1994), in biology, which can build upon a conceptual tradition of long standing (Felsenstein, 1983), and in many other areas (Abdi, 1990; Barthélemy and Guénoche, 1991). Discrete structures are commonly contrasted with, and thought to be alien from the continuous spatial structures that are used in multidimensional scaling, although 'hybrid' models have been proposed (Carroll, 1976). The purpose of this paper is to take some steps towards an integrated view of these two types of models, by showing how we can deal with the problem of fitting graphs and trees with the same formalism that is the basis of least squares methods used in multidimensional scaling (MDS).

Within the framework of least squares, discrete structural representations of proximity data are usually identified by fitting a distance matrix under constraints. For example, least squares fitting of a hierarchical tree can be done by enforcing the ultrametric inequality upon a set of non-negative quantities (Hartigan, 1967; Chandon, Lemaire and Pouget, 1980), for an additive tree we can impose the four-point condition (Sattath and Tversky, 1977; Cunningham, 1978; De Soete, 1983), and for network representations one criterion that has been used is additivity of distances along every possible path (Klauer and Carroll, 1989). Most methods used in practice, while not being least squares, are nevertheless based upon classic operations on a dissimilarity matrix to transform it into a constrained distance matrix (e.g., both the single-link or minimum method and the complete-link or maximum method for hierarchical clustering can be viewed as transformations of an arbitrary dissimilarity matrix into an ultrametric distance matrix).

By contrast, an MDS model typically represents the objects in terms of points characterized by coordinates, so that distances are not parameters to be estimated, but *functions* of other parameters. These functions may be Euclidean (as is most commonly the case) or non-Euclidean (e.g., Groenen, Mathar and Heiser, 1995). The present paper will show how to set up such an indirect parametrization, which restricts the *coordinates* to be discrete (in fact, binary), for a relatively large class of graphical structures. Following Tversky (1977)

and Shepard and Arabie (1979), we will use the concept of a *feature space*, in which each object of analysis is represented by some subset of features, while the features in turn are represented by subsets of objects. By restricting attention to models that can be formulated in terms of features, we are considering a particular subclass of graphical structures, to be called *feature graphs*.

The natural metric used in feature space is the city-block distance, which acquires several remarkable properties when the coordinates are restricted to be binary. Before discussing these in more detail, we need to introduce some notation.

2. Notation and reparametrization in terms of features

Let $O = \{o_1, \ldots, o_i, \ldots, o_n\}$ be the set of objects of analysis, and suppose that the elements of the square table $\Delta = \{\delta_{12}, \ldots, \delta_{ij}, \ldots, \delta_{nn}\}$ denote the values of a given dissimilarity function defined on the set of ordered pairs $O \times O$. We are looking for a graph representation of $\{O, \Delta\}$ by a *valued graph* (or network) $G = \{V, \mathcal{R}, \Lambda\}$, where the set $V = \{v_1, \ldots, v_i, \ldots, v_n\}$ contains the *nodes* (vertices, points), and the set $\mathcal{R}$ the *edges* (arcs, lines) of G. Thus, $\mathcal{R}$ is the collection of unordered pairs of V that defines a *relation*, that is, a subset of $V \times V$. An edge $r_{ij} = \{v_i, v_j\} \in \mathcal{R}$ is said to *join* the nodes v_i and v_j in the graph, and presence or absence of edges is indicated by the binary $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, called the *adjacency matrix*, which has $a_{ij} = 1$ if $r_{ij} \in \mathcal{R}$, and $a_{ij} = 0$ if $r_{ij} \notin \mathcal{R}$. Finally, the graph G is *valued*: we associate with each edge present ($a_{ij} = 1$) some non-negative function value λ_{ij}, called the *edge length*, collected in the matrix $\Lambda = \{\lambda_{ij}\}$, where we define $\lambda_{ij} = 0$ when $a_{ij} = 0$. A metric on G is defined by the *path-length distance*

$$d_{ij}(\mathbf{A}, \Lambda) = d(v_i, v_j) = \sum_{(i^*j^*) \in \mathcal{P}(v_i, v_j)} \lambda_{i^*j^*} , \tag{1}$$

in which $\mathcal{P}(v_i, v_j)$ is the set of edges on the *geodesic* (shortest path) between v_i and v_j. We write $d_{ij}(\mathbf{A}, \Lambda)$ because the distance depends not only on Λ, but also on $\mathbf{A}$ via the lists $\mathcal{P}(v_i, v_j)$. Thus, the path length is the sum of the edge lengths along the geodesic. If all λ_{ij} are equal, $d_{ij}(\mathbf{A}, \Lambda)$ is the usual graphical distance: a count of the number of edges in the shortest path from v_i to v_j.

Let us first consider the question of *embedding*, or realizability: under what conditions on Δ can the objects be mapped into a valued graph with some path-length distance? The answer is, that we may identify $\lambda_{ij} = \delta_{ij}$ for some subset $\mathcal{R}$, provided that δ_{ij} is positive-definite and satifies the triangle inequality $\delta_{ij} \leq \delta_{il} + \delta_{lj}$ (Hakimi and Yau, 1965). Note that symmetry is not required (if we allow two edges between any two nodes); but if δ_{ij} is in addition symmetric, it is a metric, and the result says that any metric can be embedded into a valued graph. In the presence of error, it is much to be preferred to optimize some loss function measuring the lack of fit of feasible model distances, rather than to rely on idealized conditions evaluated directly in terms of the data. Therefore, we study the *fitting* problem of finding G (in particular, some $\mathbf{A}$ and Λ) so that the least squares loss function

$$\sigma^2(\mathbf{A}, \Lambda) = \sum_i \sum_j \, (\delta_{ij} - d_{ij}(\mathbf{A}, \Lambda))^2 , \tag{2}$$

is minimal. Note that the major difficulty in (2) is finding $\mathbf{A}$, since (1) is additive in the elements of Λ, so that, once we know $\mathbf{A}$, finding Λ is just a non-negative regression problem, which can be solved by standard methods (Lawson and Hanson, 1974). How can we find out which edges to include and which to delete?

Our approach in the present paper will be to use a *reparametrization* of $d_{ij}(\mathbf{A}, \Lambda)$, which restricts attention to a certain subclass of graphs. To define the vertices of such a graph, we introduce a set of p *discrete features* $\mathcal{F} = \{F_1, \ldots, F_t, \ldots, F_p\}$. On the feature set $\mathcal{F}$ we

define a family $\mathcal{S}$ of n distinct nonempty subsets $\mathcal{S} = \{S_1, \ldots, S_i, \ldots, S_n\}$, whose union is $\mathcal{F}$. Furthermore, each feature $F_t \in \mathcal{F}$ is associated with some nonnegative *feature discriminability* parameter η_t. Every object will now be represented by some subset of features, that is, our goal will be to find a mapping $\mathcal{T}$: $o_i \in O \rightarrow S_i \in \mathcal{S}$.

To rephrase the fitting problem in terms of the mapping $\mathcal{T}$, we must have a metric on (sub)sets that parallels the path-length distance. Following Goodman (1951, 1977) and Restle (1959, 1961), we may define a metric on sets, here to be called the *feature distance*

$$d(S_i, S_j) = \mu[(S_i \cup S_j) - (S_i \cap S_j)] , \tag{3}$$

where $\mu[\cdot]$ is a measure function defined over the set of features (usually, just a count), and $A - B$ is the symmetric set difference between sets A and B. Thus the feature distance measures the extent to which S_i possesses features that S_j does not have and vice versa. By elementary means it can be shown that (3) satisfies the metric axioms, and there are a number of alternative expressions of it that enable us to naturally include the feature discriminabilities η_k, which we will consider more closely in section 3. The first and foremost property of $d(S_i, S_j)$, however, is stated in the following result.

> Theorem (Flament, 1963, p. 17).
> Let $\mathcal{L}(\mathcal{S})$ be the lattice obtained from ordering the elements of $\mathcal{S}$ by inclusion, and consider the graph representing $\mathcal{L}(\mathcal{S})$ having nodes $v_i = S_i$ and an edge between v_i and v_j whenever S_i covers S_j or vice versa. Define $d_{ij}(\mathbf{A})$ as the pathlength distance (geodesic) between nodes v_i and v_j with all edge lengths λ_{ij} equal to unity. Then the feature distance $d(S_i, S_j)$ is equal to $d_{ij}(\mathbf{A})$ in the graph representation of the lattice $\mathcal{L}(\mathcal{S})$.

If $\mathcal{S}$ is an arbitrary selection of subsets, it is understood that $\mathcal{L}(\mathcal{S})$ includes the extension of $\mathcal{S}$ with all subsets that can be formed by union and intersection of its elements. In the graphical representation of this lattice of subsets there are generally several paths from S_i to S_j, but the crucial thing is that they all have equal length; hence, they are equivalent in terms of distance. Equivalence of distinct paths follows from the fact that each edge in the graphical representation of the lattice corresponds to one single element of $\mathcal{F}$, which is the feature that distinguishes the covering subset from the covered one. While the graphical distance $d_{ij}(\mathbf{A})$ is a count taken along a path of the distinguishing features in some particular order, the feature distance $d(S_i, S_j)$ is the same count of distinguishing features, taken in any order.

Another property that turns out to be crucial for the present approach is that betweenness implies additivity: that is, if S_j is inbetween S_i and S_k in the sense that either $S_i \supset S_j \supset S_k$, or $S_i \supset S_j$ and $S_k \supset S_j$, we have $d(S_i, S_k) = d(S_i, S_j) + d(S_j, S_k)$ along the path from S_i to S_k. In this case, there need not be a direct edge from S_i to S_k. This characteristic allows us to first formulate the fitting problem (2) in terms of feature distances, next sort out the additivities in the fitted distances, and finally construct the graph by excluding edges that are sums of other edges. Thus the graphs to be constructed with the present approach will always be subgraphs of the graph representation of a lattice, which forms the embedding space of the given set of objects in much the same way as Euclidean space is used to embed a finite number of points in ordinary multidimensional scaling.

3. Introducing discriminability of features: Weighted counting

In the simplest case, the Goodman-Restle feature distance (3) is a straight count of the features in the set difference between the union and the intersection of two subsets. Although there are a number of interesting re-expressions of the feature distance in terms of set operations, what we need for our MDS-like fitting problem is an expression in terms of

coordinates. Let $\mathbf{E} = \{e_{it}\}$ be a binary matrix of order $n \times p$, which indicates which features of $\mathcal{F}$ are included in each of the n subsets in $\mathcal{S}$ that represent the objects. Since $\mathbf{E}$ characterizes objects as subsets of features (but also features as subsets of objects), it is (the transpose of) a *point-set incidence matrix* (see Roberts, 1976, page 60). When $\mu[\cdot]$ is just a counting measure, (3) becomes

$$\begin{aligned} d(S_i, S_j) &= \sum_t \{(1 - e_{it})e_{jt} + (1 - e_{jt})e_{it}\} \\ &= \sum_t \{e_{jt} - e_{it}e_{jt} + e_{it} - e_{jt}e_{it}\} \\ &= \sum_t (e_{it} - e_{jt})^2 = \sum_t |e_{it} - e_{jt}| , \end{aligned} \tag{4}$$

where the last equality follows from the binary nature of $\mathbf{E}$. Thus, the feature distance is equal to a city-block metric on a space with binary coordinates, a metric better known as the *Hamming distance*. This distance is commonly used as a dissimilarity coefficient, in a situation where the e_{it} are presence-absence data, or as a theoretical device (Boorman and Arabie, 1972), but – to the best of the author's knowledge – it has never been used as a structural model to be fitted to dissimilarity data.

Especially for fitting purposes, it is useful to take one further step and to go from a simple count to a weighted count, that is, to generalize (4) into

$$d_{ij}(\mathbf{B}) = \sum_t \eta_t |e_{it} - e_{jt}| = \sum_t |b_{it} - b_{jt}| , \tag{5}$$

where the b_{it}'s are still binary, albeit not necessarily (0,1) variables, collected in the $n \times p$ matrix $\mathbf{B} = \{b_{it} = \eta_t e_{it}\}$, and where the discriminabilities η_t are nonnegative parameters to be estimated. Thus, the weighted feature distance defined in (5) allows for a differential contribution of the features to the overall length of the path from S_i to S_j. In geometrical terms, the introduction of feature discriminabilities turns the hypercube corresponding to $\mathbf{E}$ into a rectangular parallelepiped corresponding to $\mathbf{B}$.

It can be shown that the Theorem stated in the previous section still holds for the weighted feature distance, if it is adjusted to allow for unequal λ_{ij}. Since each edge in the graphical representation of the lattice $\mathcal{L}(\mathcal{S})$ corresponds to one feature in $\mathcal{F}$, we can associate exactly one feature discriminability η_t in (5) with each edge length λ_{ij} in (1). For example, if the set of edges on the shortest path between v_i and v_j would be $\mathcal{P}(v_i, v_j) = \{(v_i, v_k), (v_k, v_l), (v_l, v_j)\}$, there will be three features F_1, F_2, and F_3 on which S_i and S_j are different, with the edge lengths being related by the one-to-one mapping $\lambda_{ik} = \eta_1$, $\lambda_{kl} = \eta_2$, and $\lambda_{lj} = \eta_3$. Hence we have $d_{ij}(\mathbf{A}, \Lambda) = \lambda_{ik} + \lambda_{kl} + \lambda_{lj} = \eta_1 + \eta_2 + \eta_3 = d_{ij}(\mathbf{B})$ in this example.

4. Algorithm for fitting a feature graph

Due to the pioneering work of Hartigan (1967), Cunningham (1978), Arabie and Carroll (1980), De Soete (1983), Mirkin (1987), and others, least squares fitting of discrete models is gradually gaining ground over various *ad hoc* procedures that were once more common. After replacement of the path-length distance $d_{ij}(\mathbf{A}, \Lambda)$ by the feature distance $d_{ij}(\mathbf{B})$, the least squares loss function (2) that we are interested in becomes

$$\sigma^2(\mathbf{B}) = \sum_i \sum_j (\delta_{ij} - d_{ij}(\mathbf{B}))^2 , \tag{6}$$

which must be minimized over all binary valued matrices $\mathbf{B}$. Because the feature distance is additive over features, it is possible to employ an alternating least squares (ALS) scheme, fitting the model one feature at a time, given some starting values $\{\underline{b}_{it}\}$. Explicitly, given

the current values $\{\underline{b}_{is}\}$ for $s \neq t$, $\underline{\delta}_{ij}$ is defined as $\underline{\delta}_{ij} = \delta_{ij} - \sum_{s \neq t} |\underline{b}_{is} - \underline{b}_{js}|$, the original dissimilarity corrected for the contribution of the fixed variables. Substituting (5) into (6), and inserting $\underline{\delta}_{ij}$, we find that the ALS subproblem for feature t is to minimize, given $\underline{\delta}_{ij}$,

$$\sigma^2(\mathbf{b}_t) = \sum_i \sum_j \, (\underline{\delta}_{ij} - |b_{it} - b_{jt}|)^2 \, , \tag{7}$$

over the binary n-vector $\mathbf{b}_t$. This minimization subtask is a one-dimensional MDS problem with the coordinates restricted to form a bipartition, and therefore the cluster differences scaling (CDS) algorithm of Heiser and Groenen (1997) applies, with number of clusters equal to two. The ALS algorithm cycles over CDS subtasks until convergence.

Let us have a closer look at this particular CDS substask, by resolving **B** again in its discrete and continuous factors. Writing $|b_{it} - b_{jt}| = \eta_t \{(1 - e_{it})e_{jt} + (1 - e_{jt})e_{it}\}$, setting the partial derivative of (7) with respect to η_t equal to zero and simplifying, shows that, for any given bipartition $\{e_{it} \mid i = 1, \ldots, n\}$ the optimal value of the discriminability parameter for feature t is equal to $\max(0, \hat{\eta}_t)$, with $\hat{\eta}_t$ denoting the unconstrained minimizer

$$\hat{\eta}_t = \frac{1}{n_t(n - n_t)} \sum_i \sum_j \, e_{it}(1 - e_{jt})\underline{\delta}_{ij} \, , \tag{8}$$

where n_t is the number of objects in one group, and $n - n_t$ the number of objects in the other. Thus, the length of edge t in the fitted feature graph will be equal to the average corrected dissimilarity between the two groups of objects that constitute that particular feature. If the features are exclusive, $e_{it}(1 - e_{jt})\underline{\delta}_{ij} = e_{it}(1 - e_{jt})\delta_{ij}$, and (8) becomes just the average between-group dissimilarity. If the features are not exclusive and $\hat{\eta}_t$ is relatively large, then the corresponding bipartition must be a good discriminator by itself, on top of the contribution of the other features, since we always have $\underline{\delta}_{ij} \leq \delta_{ij}$; this justifies the name discriminability parameter.

We still have to indicate how to find **E**. Loss function (7) is quadratic in one column (size n) of the binary matrix **E**, so this subtask still is a hard combinatorial problem, even though its size is reduced by a factor p with repect to loss function (6). The present implementation uses a nesting of several random starts (within features and across features), together with K-means type reallocations. Heiser and Groenen (1997) have described a strategy called *Fuzzy Steps* to alleviate the local minimum problem for CDS, but it looks like the problem here is especially difficult across features (in the ALS phase), not so much within features. A more extended discussion of the algorithmic aspects of finding **E** is in preparation.

5. Example: Henley's (1969) animal terms

To see how the feature graph procedure works, dissimilarity data originally collected by Henley (1969) will be analyzed as an example. In a psychological experiment on semantic memory structure, 21 subjects were instructed to freely list from memory any animal terms they knew. From the total set of animal terms mentioned, 12 common ones were selected. Dissimilarities were computed for each pair of terms as the average (across subjects) of the proportion of items separating them in each list. Figure 1 displays an additive tree representation of the animal terms, given by Abdi (1990), with a percentage of variance accounted for of 73.0%. Many other tree representations for this example can be found in Barthélemy and Guénoche (1991).

The feature-graph method was applied with the number of features ranging from 1 to 10. The one-feature solution, which is a simple two-cluster split, had a prevalence of 7 out of 10 random initial bipartitions, and accounted for 21.9% of the variance. It splits {lion, bear, pig, sheep, goat, cow, horse} from {cat, dog, mouse, rabbit, deer}. Note that this split cannot be obtained by cutting anyone of the edges of the optimal tree, while it does

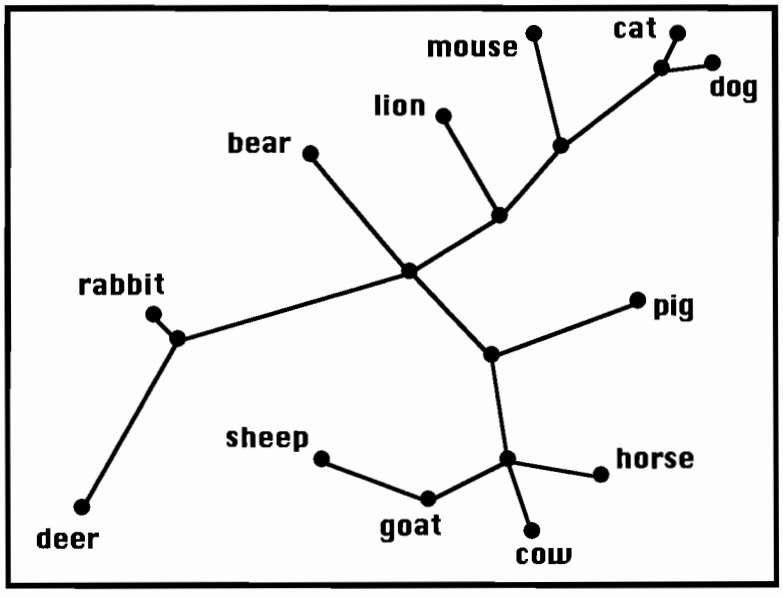

Fig. 1: Additive tree for Henley's animal terms

seem to represent the best split in terms of the dissimilarity data. The percentage of variance accounted for (VAF) for all ten solutions is given in Table 1. This table also gives for each solution the DAF (percentage of Dispersion Accounted For), defined as the sum of squared fitted distances divided by the sum of squared dissimilarities. DAF is the scale-free goodness-of-fit measure that is maximized when the badness-of-fit measure (6) is minimized.

Table 1. Goodness of fit for feature graph representations of the Henley (1969) data

# features:	1	2	3	4	5	6	7	8	9	10*
DAF	63.3	83.4	89.4	93.9	95.8	96.9	97.6	98.1	98.5	97.3
VAF	21.9	41.9	54.0	71.9	79.2	82.3	84.7	87.7	90.3	83.9

*solution with 4 unicities

We see from Table 1 that a VAF just above the percentage of the tree solution is reached with the five-feature solution (79.2%), which has a DAF of 95.8%. This solution, which does not yet discriminate all objects from each other (leaving seven objects in three small clusters), is shown in Figure 2. While the terms in the clusters {cat, dog} and {goat, cow, horse} in the feature graph are also close together in the additive tree in Figure 1, this is not the case for the cluster {bear, pig}. Another major difference is that the feature graph is not tree-like at all. As to the interpretation of the five-feature solution in Figure 2, it is clear that {deer} is an isolate (there is one feature that contrasts it with all other terms, with a discriminability of approximately 20), and that there is a "domestic" versus "wildlife" feature contrasting, from top to bottom, {sheep, cow, horse, goat, cat, dog} from {pig, bear, lion, mouse, rabbit, deer}, with a discriminability of about 14. A third important split is {cat, dog, lion, mouse} versus the rest, with discriminability 16.

A more differentiated representation arises if we look at one of the solutions with a higher number of features. How many features to take is not only a matter of amount of fit that is

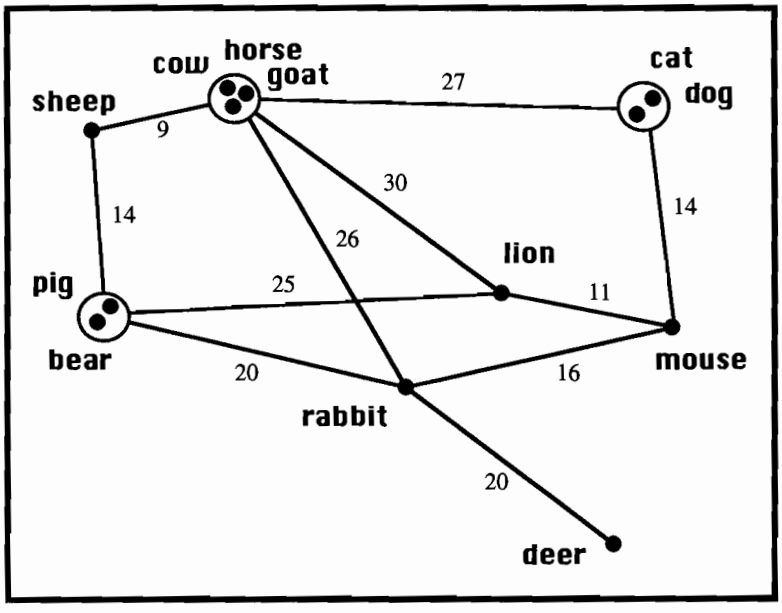

Fig. 2: Five-feature solution of Henley's data (with estimated edge lengths)

deemed acceptable, but also depends on the issue of how many edges need to be kept, or conversely, how many additivities there are in the fitted distances. Judged by the number of edges needed, while still accounting for a reasonable amount of variance, a special ten-feature solution was selected as the best one (see the last column of Table 1; its graph with 24 edges is displayed in Figure 3). It consists of six common features (i.e., features shared by more than one object) and four unique features (not shared by any other object).

Figure 3 contains two types of nodes: the closed circles, which represent the objects of analysis, and the open circles, called *latent nodes*, which represent subsets of features that can be obtained by taking either the union or the intersection of the feature subsets characterizing two other nodes. Remember that the fitted feature graph is a subgraph of the

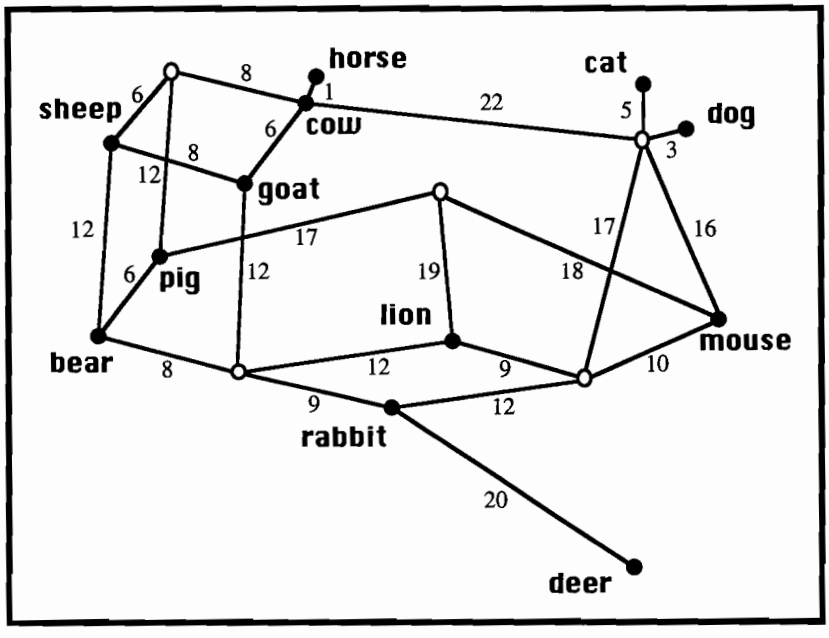

Fig. 3: Ten-feature solution of Henley's data, with 6 common features and 4 unique ones (open circles are latent nodes)

graph representation of the lattice of feature subsets, and latent nodes are other elements of this lattice that can be included afterwards, to make the graph simpler in terms of its pathways and number of edges. As a good example of the effect of the introduction of a latent node, consider four objects characterized by the subsets (BCD), (ACD), (ABD), and (ABC), and assume equal discriminability of the features. Then all distances are equal, and the objects are mapped as four points on a regular tetrahedron, with six edges. Introducing the latent node (ABCD), which is the union of each of the pairs of subsets, allows us to construct a star graph, in which there are only four edges, one between each of the manifest nodes and the latent node, and no one among the manifest nodes themselves.

The fitted edge lengths are also given in Figures 2 and 3 (rounded to integer numbers). An edge is not included in the graph if its length is the sum of two other edge lengths (a rather simple algorithm looping over all triads is sufficient to sort this out). To reconstruct the distance between two terms (and hence their dissimilarity), we just have to add the edge lengths along the shortest path between them. It will be noted that there are several instances of distinct paths with equal length. Comparing the two feature-graph solutions, it appears that there are primarily local changes: one is slightly more (less) differentiated than the other, a result that makes sense.

6. Some special cases

Modeling considerations can be formulated in terms of the lattice of subsets $\mathcal{L}(\mathcal{S})$ as follows: given Δ, or some approximation of it satisfying the triangle inequality, does there exist feature distances $d(S_i, S_j)$ that arise from the feature graph of a family of subsets that has a certain property? Consequently, the fitting problem would become one of optimizing loss function (6) over a family of feature sets that have a specified structural property. In this section, we will briefly indicate some examples of structural properties that may be handled in the present framework; a more detailed treatment of them is in preparation.

Before considering the special cases, however, we discuss a technical issue that needs to be settled first. In a feature distance model, the role of presence and absence is symmetric; that is, we can always replace all elements e_{it} of a whole column of $\mathbf{E}$ by their complement $1 - e_{it}$ without changing the distances. To illustrate, the two weighted binary matrices

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 & 0 & 2 & 0 & 1/2 & 0 & 3/2 \\ 0 & 1 & 2 & 0 & 0 & 1/2 & 0 & 3/2 \\ 0 & 1 & 0 & 2 & 1/2 & 0 & 0 & 3/2 \\ 0 & 1 & 0 & 2 & 0 & 1/2 & 3/2 & 0 \end{bmatrix}$$

generate the same feature distances among their rows. Thus we can freely add the complements of any column of the incidence matrix, provided that we *half* the corresponding discriminabilites. Any $n \times 2$ matrix formed by concatenating some column of $\mathbf{E}$ with its complement has the property that it has row sums equal to one, and such a matrix is called the *indicator matrix* of a feature.

Now suppose that the features are *nested*: that is, if $\mathbf{G}_t$ is the indicator matrix of feature F_t and $\mathbf{G}_s$ is the indicator matrix of feature F_s, than the matrix $\mathbf{G}_t'\mathbf{G}_s$ has at least one element equal to zero. Nestedness implies that one feature separates a subgroup from one of the two groups formed by the other feature. For instance, the bipartitions {(ABCD), (EFG)} and {(EF), (ABCDG)} are nested, since (EF) is a subset of (EFG) and (ABCD) is a subset of (ABCDG). Then, by a famous result of Buneman's (1971), the feature distance satisfies the four-point property that characterizes *additive trees* if and only if all its features are nested. Additive trees thus form an important special case of feature graphs, in which each edge corresponds to exactly one feature (or *split*).

The case of a *linear array* (Goodman, 1951, 1977; Restle, 1959, 1961), called *Guttman scale* in psychometrics, is obtained when the features are not only nested, but have an additional property. In terms of the feature incidence matrix **E**, this property implies that each column of **E** consists of either a single run of zeros followed by a single run of ones or of a single run of ones followed by a single run of zeros. When $n - 1$ distinct features have this structure, the feature graph has $n - 1$ edges, connecting the objects in a certain order, and no latent nodes. Except for the two endpoints, which have degree one, all nodes have degree two. For an exact characterization of the Guttman scale, see Holman (1995).

A *hierarchical tree* is a rooted additive tree with the extra requirement that the distance from any endpoint to the root is equal. In a feature graph, the root corresponds to the latent node that has all features, that is, $\mathcal{F}$. Then the first feature defines the first split in two groups of objects, the second feature splits one of these groups further down into subgroups, and so on. So the features are again nested. The hierarchical tree is a more parsimonious model than the additive tree, because the requirement of equal distance to the root puts restrictions on the discriminabilities. The characterization of trees in terms of a feature model is due to Tversky (1977).

The last example of a family of subsets that satisfies a specific structural property is the *circumplex* or *radex* (Guttman, 1954). It is characterized by the *circular ones* property, which implies that each column of **E** consists of either a single run of zeros bordered by a run of ones on one or both sides, or of a single run of ones bordered by one or two run(s) of zeros. The graph of a regular circumplex is like a closed simple chain, with exactly n edges, if each feature divides the objects in equal groups (when n is even). When divisions in unequally sized groups are included in the feature set, the graph of a circumplex becomes more complicated. In the complete case, it looks like a network spanned over a (half)sphere (Heiser, 1981, Chapter 4).

7. Discussion

We have seen that a metric defined on the symmetric set difference between sets of features can be used as a general framework for fitting a particular class of graphs, which includes additive trees, hierarchical trees and circumplex structures. It was shown that we can find out which edges to include in the graph by formulating the problem in terms of a lattice of subsets, using a weighted count of feature differences (the feature discriminabilities). The algorithm presented, based on alternating least squares and on cluster differences scaling, is still in its early stage of development. It always converges to a local minimum, but as is usual in this type of problem, there are an awful lot of local minima. On the positive side, it is the first systematic method to fit the Hamming distance.

A crucial ingredient of this approach to finding graph representations is the fact that inclusion relations between feature sets lead to additivity of distance along paths in a graph. In fact, Hutchinson (1989) and Klauer and Carroll (1989) used the criterion of dropping direct edges by looking at additivity of link length as their main graph construction strategy. But they applied this criterion to the dissimilarities, rather than to the fitted distances, as is proposed here. Feature graphs are similar to Corter and Tversky's (1986) extended similarity trees, but exactly how these two models are related needs further study. In any case, it seems clear that additive trees and other restricted representations do not show up spontaneously in real examples, although the method reproduces a circumplex, for example, when the data are error-free.

Choosing the number of features p is a matter that requires experience and cannot be settled yet with clear-cut rules. In most examples analyzed so far, the number of features needed to get good fit is in the neighborhood of $n/2$. Also, it appears that, as soon as p is in the

range of values where the fit stabilizes, solutions with one feature more or one feature less are only different in their fine structure, as was the case in the example of the Henley (1969) data. There is a trade-off to be made with the number of links in the graph, a quantity that increases nonlinearly with p, and which we want to have as small as possible. Making a good trade-off is complicated by the fact that we can often reduce the number of links by including latent nodes, without it being clear how to do this optimally.

Unlike methods based on distance constraints, feature graph fitting can be extended without too much trouble to well-known variants of MDS, such as individual differences scaling (INDSCAL) and two-mode scaling (unfolding), possibly combined with nonlinear transformations of the data. An easy way to recognize this flexibility is to view the basic distance model (5) as a squared Euclidean distance (since the deviations $e_{it} - e_{jt}$ are zero or (minus) one, we just have to reparametrize η_t as the square of some other non-negative parameter). Then the feature graph loss function (6) is identical to Takane, Young and De Leeuw's (1977) SSTRESS loss function with restrictions on the configuration.

References:

Abdi, H. (1990): Additive tree representations, In: *Trees and Hierarchical Structures*, Dress, A. et al. (Eds.), 43-59, Springer Verlag, Berlin.

Arabie, P., and Carroll, J.D. (1980): MAPCLUS: A mathematical programming approach to fitting the ADCLUS model, *Psychometrika,* **45**, 211-235.

Arabie, P., and Hubert, L. (1992): Combinatorial data analysis, *Annual Review of Psychology,* **43**, 169-203.

Barthélemy, J.-P. and Guénoche, A. (1991): *Trees and Proximity Representations*, Wiley, New York.

Boorman, S.A. and Arabie, P. (1972): Structural measures and the method of sorting, In: *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Shepard, R.N. et al. (Eds.), 225-249, Seminar Press, New York.

Buneman, P. (1971): The recovery of trees from measures of dissimilarity, In: *Mathematics in the Archaeological and Historical Sciences,* Hodson, F.R. et al. (Eds.), 387-395, Edinburgh University Press, Edinburgh.

Carroll, J.D. (1976): Spatial, non-spatial and hybrid models for scaling, *Psychometrika,* **41**, 439-463.

Chandon, J.L., Lemaire, J., and Pouget, J. (1980): Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, *R.A.I.R.O. Recherche Opérationelle,* **14**, 157-170.

Corter, J.E., and Tversky, A. (1986): Extended similarity trees, *Psychometrika*, **51**, 429-451.

Cunningham, J.P. (1978): Free trees and bidirectional trees as representations of psychological distance, *Journal of Mathematical Psychology,* **17**, 165-188.

De Soete, G. (1983): A least squares algorithm for fitting additive trees to proximity data, *Psychometrika,* **48**, 621-626.

Felsenstein, J. (Ed.)(1983): *Numerical Taxonomy*, Springer Verlag, Heidelberg.

Flament, C. (1963): *Applications of Graph Theory to Group Structure*, Prentice-Hall, Englewood Cliffs, New Jersey.

Goodman, N. (1951): *The Structure of Appearence*, Bobbs-Merrill, Indianapolis, Indiana.

Goodman, N. (1977): *The Structure of Appearence* (3rd ed.), Reidel, Dordrecht, Holland.

Groenen, P.J.F., Mathar, R., and Heiser, W.J. (1995): The majorization approach to multidimensional scaling for Minkowski distances, *Journal of Classification*, **12**, 3-19.

Guttman, L. (1954): A new approach to factor analysis: The radex, In: *Mathematical thinking in the social sciences*, Lazarsfeld, P.F. (Ed.), 258-348, The Free Press, Glencoe, Illinois.

Hakimi, S.L., and Yau, S.S. (1965): Distance matrix of a graph and its realizability, *Quarterly of Applied Mathematics*, **22**, 305-317.

Hartigan, J.A. (1967): Representation of similarity matrices by trees, *Journal of the American Statistical Association,* **62**, 1140-1158.

Heiser, W.J. (1981): *Unfolding analysis of proximity data*, Unpublished doctoral dissertation, University of Leiden, The Netherlands.

Heiser, W.J., and Groenen, P.J.F. (1997): Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima, *Psychometrika*, **62**, in press.

Henley, N.M. (1969): A psychological study of the semantics of animal terms, *Journal of Verbal Learning and Verbal Behavior*, **8**, 176-184.

Holman, E.W. (1995): Axioms for Guttman scales with unknown polarity, *Journal of Mathematical Psychology*, **39**, 400-402.

Hutchinson, J.W. (1989): NETSCAL: A network scaling algorithm for nonsymmetric proximity data, *Psychometrika,* **54**, 25-52.

Klauer, K.C. (1994): Representing proximities by network models, In: *New Approaches in Classification and Data Analysis*, Diday, E. et al. (eds.), 493-501, Springer Verlag, Heidelberg.

Klauer, K.C., and Carroll, J.D. (1989): A mathematical programming approach to fitting general graphs, *Journal of Classification,* **6**, 247-270.

Lawson, C.L., and Hanson, R.J. (1974): *Solving least squares problems*, Prentice Hall, Englewood Cliffs, NJ.

Mirkin, B.G. (1987): Additive clustering and qualitative factor analysis methods for similarity matrices, *Journal of Classification*, **4**, 7-31.

Restle, F. (1959): A metric and an ordering on sets, *Psychometrika,* **24**, 207-220.

Restle, F. (1961): *Psychology of Judgment and Choice*, Wiley, New York.

Roberts, F.S. (1976): *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*, Prentice Hall, Englewood Cliffs, New Jersey.

Sattath, S., and Tversky, A. (1977): Additive similarity trees, *Psychometrika*, **42**, 319-345.

Shepard, R.N., and Arabie, P. (1979): Additive clustering: Representation of similarities as combinations of discrete overlapping properties, *Psychological Review*, **86**, 87-123.

Takane, Y., Young, F.W., and De Leeuw, J. (1977): Nonmetric individual differences in multidimensional scaling: An alternating least quares method with optimal scaling features, *Psychometrika*, **42**, 7-67.

Tversky, A. (1977): Features of similarity, *Psychological Review*, **84**, 327-352.