Semiparametric estimation of (constrained) ultrametric trees

Wedel, Michel; DeSarbo, Wayne S.

Semiparametric Estimation of (Constrained) Ultrametric Trees

Michel Wedel*
University of Groningen

Wayne S. DeSarbo
Pennsylvania State University

SOM theme B: Marketing and Networks

November 1996

Abstract

This paper is concerned with the semi-parametric estimation of Ultrametric tree-representations of subjects' paired comparisons of stimuli, and captures subject heterogeneity using a finite mixture formulation. In many other approaches to the analysis of subjects decision processes, such finite mixture models have been gainfully applied. A new likelihood based estimation methodology is presented for Ultrametric tree structures that accommodates the Ultrametric constraints. This estimation procedure in addition permits the incorporation of a variety of additional external restrictions on the tree structure. Correlations among the observed dissimilarities are allowed for. The performance of the method to identify Ultrametric trees is investigated on synthetic data and an empirical application to published data from Schiffman, Reynolds, and Young (1981) is provided. The ability to deal with specific constraints on the tree-topology is demonstrated.

Key words: Hierarchical Clustering, Semi-parametric Estimation, Finite Mixtures, UltrametricTrees, Common Features Model, Respondent Heterogeneity.

1. Introduction -Tree Models of Dissimilarity

In the past two decades, the psychometric literature has seen much interest in tree-structure models of stimulus paired comparison processes. Models that have been proposed include traditional hierarchical clustering methods (cf. Johnson, 1967), Ultrametric trees (cf. Carroll, 1976), additive trees (cf. Sattah & Tversky, 1977), multiple trees (cf. Carroll & Pruzansky 1975), and extended similarity trees (cf. Corter & Tversky 1986), amongst others. Extensive reviews of these and other developments has been provided by DeSarbo, Manrai, and Manrai (1993) and Corter (1996). In this paper, we are concerned with capturing heterogeneity in subjects decision processes involved in paired comparisons of stimuli. This heterogeneity is represented by a finite mixture of several Ultrametric tree structures with different topologies. We present a semi-parametric estimation methodology that accommodates the Ultrametric constraints in the estimation process. This estimation procedure allows us, in addition, to optionally incorporate a variety of external restrictions in the fitted tree structure. We start by briefly reviewing the extant literature. For a more detailed discussion of previous work, we refer to DeSarbo, Manrai, and Manrai (1994) and Corter (1996).

Tree models of dissimilarity judgements represent a set of stimuli as nodes in a connected, undirected graph. The nodes are connected by arcs, where cycles do not occur. The proximity between the stimuli is represented in the tree by the distance, or path-length, between the nodes in the graph. The two dominant types of tree structures heeded in the psychometric literature are Ultrametric and the additive trees, which are characterised by the Ultrametric inequality on all triples of stimuli, and, respectively, the additive inequality defined on all quadruples of stimuli (cf. Corter, 1996). Whereas Ultrametric trees have a unique root, additive trees do not. In this manuscript we restrict ourselves to Ultrametric tree representations of similarity. One way to interpret Ultrametric tree structures is in terms of a common features model of similarity (Tversky, 1977) in which the feature sets have a hierarchical structure. Here, the sum of the lengths of the arcs emanating from the root of the tree and leading to the least common ancestor node of two particular stimuli is a measure of the features shared by these two stimuli (Corter & Tversky, 1986).

1.1 Heterogeneity

In many psychometric models of human decision behaviour, attempts have been made to account for heterogeneity of the decision process among subjects. In recent years one relatively popular way of modelling heterogeneity has been through the use of finite mixture models. For example, finite mixture formulations have been used in spatial models describing outcomes of the decision process (cf. DeSarbo, Manrai& Manrai, 1994, Wedel & DeSarbo, 1995), and in (generalized) linear models describing compensatory decision processes (cf. Wedel & DeSarbo 1994, Wedel & DeSarbo, 1995, for reviews). However, relatively few studies in the psychometric and related literatures have dealt with perceptual heterogeneity in tree-structure models. Rao and Sabavala (1981) suggest performing a hierarchical clustering of stimuli for each of a number of classes (defined a-priori or derived post hoc by clustering the subjects). Their approach has the disadvantages of being based on heuristics, maximizing two unrelated criteria in the two subsequent steps of the procedure. Carroll, Clark and DeSarbo (1984) propose a tree model that posits a common Ultrametric tree topology for each of a number of subjects or classes of subjects (derived prior to the analysis), and allows the different classes to have different branch lengths. However, this model is unable to capture major forms of structural heterogeneity, where different Ultrametric tree topologies among classes of subjects are involved (see below). Carroll and Pruzansky (1980) propose multiple tree structures in which observed dissimilarities are represented by a sum of hierarchical trees that each satisfy the Ultrametric inequality. In their model multiple trees are added to form a representation of dissimilarity. Their approach does not explicitly deal with subject heterogeneity, however. Somewhat surprisingly, perhaps, a mixture model extension of the multiple trees approach has not yet appeared in the literature. The purpose of this paper is to fill this gap. The mixture model identifies Ultrametric trees, class sizes and membership, all simultaneously. In addition, the tree topology is not restricted to be the same across classes. As an example, we will demonstrate in our empirical analysis of paired comparisons of sensory stimuli, that a sample of subjects groups into two clusters of equal size. Each of the two classes of subjects is well represented by an Ultrametric tree structure with a specific topology (see Figure 5 below). Without prior knowledge of class membership, an aggregate analysis of the entire sample produces an Ultrametric tree that represents the structure of neither class and produces a poor fit to the data (see Figure 4), i.e. the true structure is masked by respondent heterogeneity and the inability of aggregate level analyses to depict such differences among classes.

1.2 Estimation Algorithms

Early estimation of tree models has involved the use of hierarchical clustering procedures. However, there are several disadvantages related to the use of these methods. First, hierarchical clustering methods are mere heuristics for identifying a single Ultrametric tree structure from paired comparisons data. Consequently, different algorithms (e.g. single, complete, average linkage) often lead to different solutions and there is little theory to choose among them. In addition, several of these procedures are burdened with problems such as chaining, non-uniqueness, and inversions (Morgan & Ray, 1995). Alternative procedures have been developed that focus on the estimation of the tree-structure by minimizing some statistical criterion of fit. Hartigan (1967) minimizes a least squares function between the observed dissimilarities and fitted distances by performing local operations on the tree. Due to its combinatorial nature, however, this type of algorithm is computational rather intensive and can be applied only to smaller data sets. Carroll and Pruzansky (1980) were the first to propose a mathematical programming approach for ultrametric tree estimation, that minimizes a least-squares criterion. Their procedure adds a penalty, which measures the departure from the ultrametric inequality, to the least-squares criterion function, and utilizes a steepest descent gradient search to estimate the ultrametric distances. DeSoete (1983) uses a computational more efficient penalty function, employing an exact sequential unconstrained minimization procedure, and the numerically more stable conjugate gradient nonlinear minimization method to estimate the tree distances, minimizing a least squares function of fit. These studies have focussed on deterministically estimating the distances under the Ultrametric constraints, but the stochastic nature of the respondents decision process, interdependencies among the dissimilarity judgements, as well as additional external constraints have not been dealt with.

These traditional methods of deriving Ultrametric trees have assumed a deterministic model underlying the proximity judgements: $d_{ijn} = \delta_{ijn} + e_{ijn}$, where I, j index stimuli, n indicates subjects, d_{ijn} are the observed proximity judgements, δ_{ijn} the Ultrametric tree distance, and e_{ijn} an error term. Here e_{ijn} has typically not been assumed to have an explicit distributional form, but is considered to be an approximation error. Yet, as in much of the psychologic literature on consumer decision processes, subjects response process may be considered stochastic, influenced by unobserved variables such as fatigue or loss of attention. Correlations among the dissimilarities may arise due to their being provided by the same individual, or due to the order in which the stimuli are presented. Using a least squares function of fit implicitly assumes the dissimilarities uncorrelated, which may negatively affect the accuracy of the estimated tree structure.

1.3 Constraints on the Tree Topology

When fitting ultrametric trees to dissimilarity data, external constraints on the tree topology may be derived from prior theory on the structure of the stimulus set, for example in terms of known features. They arise, for example, in market structuring (Allenby 1989), or in spatial contexts from contiguity constraints (Gordon 1973). Gordon (1980) proposes amongst others a two algorithms for constrained clustering: a divisive method that sequentially divides the sample into finer groups, and a dynamic programming algorithm. Other approaches to constrained classification have been described by for example Ferligoj and Batagelj (1982) and DeSarbo and Mahajan (1984). In this paper, we use a sequential quadratic programming algorithm to estimate the ultrametric distances, while optionally imposing user specified external restrictions. Different restrictions may be imposed on the tree topology for different classes, so that subjects can be classified a posteriori according to potentially rival hypotheses on the structure of their feature representation of the stimulus set.

Below, we describe the semi-parametric finite mixture model and the constrained estimation procedure. We demonstrate its performance by analysing synthetic data, and provide an empirical application to published data from Schiffman, Reynolds and Young (1981). There, we demonstrate its ability to deal with correlations among the dissimilarity measures and class specific constraints on the tree-topology.

2. The Semiparametric Ultrametric Tree Model

Let us first establish the notation. We let n denote subjects, I, j, k denote stimuli, and s denote classes of subjects. The data: d_{nij} presents the observed dissimilarity of stimulus I and stimulus j by subject n. We assume S unobserved classes in proportions $\pi_{s.}$ Given class s, we assume the p=I(I-1)/2 dissimilarities for subject n to follow a multivariate normal distribution:

$$\phi_s(\boldsymbol{D}_n \,|\, \boldsymbol{\Delta}_s, \boldsymbol{\Sigma}_s) \; = \; (2\,\pi)^{-p/2} \,|\boldsymbol{\Sigma}_s|^{-1/2} \mathrm{exp}\left[-1/2\,(\boldsymbol{D}_n - \boldsymbol{\Delta}_s)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{D}_n - \boldsymbol{\Delta}_s)\right] \tag{1}$$

where $\Delta_s = ((\delta_{ijs}))$ is the expected value of $\mathbf{D}_n = ((d_{ni}))$ given class s, and Σ_s its (pxp) covariance matrix in class s. In many applications, it is appropriate to assume a log-normal distribution for the distances. This constrains the fitted distances to be positive, which is logically consistent. Such a log-normal distribution may be employed by transforming the measured dissimilarities by natural logarithms and using the multivariate normal density in (1) for the log-dissimilarities, $\ln(d_{ijn})$. The multivariate normal density allows one to model the covariance structure of the paired comparisons within derived classes.

Note that in the previous work on ultrametric tree estimation, the dissimilarities have been assumed independent, so that least squares estimation can be employed. However, dissimilarities may be correlated because they are provided by the same individual, or correlated due to order effects. We denote the p(p-1)/2 correlation matrix of the judgements of pairs (i,j) and (k,l) in class s by $P_s = ((\rho_{ij,kl,s}))$. Since the number of parameters in the unrestricted covariance matrix that needs to be estimated is very large in the full model, a more parsimonious representation of the covariance matrix is called for. For example, if all the variances of the pairs for a particular class are assumed equal, $\sigma_{ijs}{}^2=\sigma_s{}^2$, the covariance matrix can be written as: $\Sigma_s = \sigma_s{}^2 \; P_s$. For example, possible parametrizations of the correlation matrix arise by considering the order in which the judgements are provided:

1. Diagonal correlation matrix: $\rho_{ij,kl,s} = 0$ for all (i,j) = (k,l); such a correlation matrix arises when the dissimilarity judgements obey a random walk process with constant variance.
2. Equicorrelation matrix: $\rho_{ij,kl,s} = \rho_s$, for all pairs (i,j) and (k,l); such a correlation matrix arises when all dissimilarities provided by the same subject are equally correlated, irrespective of the serial position of the pairs.
3. Serial correlation matrix: $\rho_{ij,kl,s} = \rho_s{}^{|q|}$, with q the difference in serial positions of pairs (i,j) and (k,l); such a correlation structure arises when the dissimilarity judgements display an AR(1) process.

For each class s, it is assumed that each triple of expected distances satisfies the ultrametric inequality:

$$\delta_{ijs} \leq \max(\delta_{iks},\delta_{jks}), \quad \forall(i,j,k,s) \tag{2}$$

The set of constraints in (2) imposes the ultrametric inequality for each class s separately, thus allowing the tree topology to differ by class. In addition, the estimation methodology detailed below allows one to impose a variety of external constraints on the ultrametric tree solutions for each of the classes. Specifically, sets of stimuli can be constrained to be in the same branch of the ultrametric tree. For example, if two stimuli, I and j, are constrained to be joined at the lowest node of the tree in class s, then $\delta_{iks} = \delta_{jks}, \ \forall \ k \neq i,j$. In general, let a subset of stimuli, pairs of which being denoted by I and j, be joined in one branch of the tree in class s, (I, j) $\in$ C_s, then:

$$\delta_{iks} = \delta_{jks}, \quad \forall \ (i \neq j \in C_s \wedge k \notin C_s) \tag{3}$$

Each such a restriction partitions the full tree for the I stimuli in that class, into two subtrees, of say I_1 and I_2 stimuli (I_1 + I_2 = I). More than one restriction of the type (3) may be imposed for each class, and different restrictions may imposed on the trees in different classes, including those discussed in Gordon (1980) and Ferligoj and Batageli (1982).

2.1 Estimation

The likelihood:

$$L(\mathbf{\Delta},\mathbf{\Sigma} \mid \boldsymbol{D}) = \prod_{n=1}^{N} \sum_{s=1}^{S} \pi_s \phi_s(\boldsymbol{D}_n \mid \mathbf{\Delta}_s,\mathbf{\Sigma}_s) \tag{4}$$

is maximized under the ultrametric constraints on the fitted distances provided by (2) using an E-M algorithm (Dempster, Laird & Rubin, 1977), implementing the Sequential Quadratic Programming (SQP) method (cf. Jamshidian & Bentler, 1993) in the M-step to enforce the active constraints. Here, Δ=(Δ_s), Σ=(Σ_s), and **D**=(**D**$_n$) .The algorithm maximizes the likelihood in a series of E-M major iterations, and minor SQP iterations within each M-step. The SQP method operates as follows. If we collect all parameters at minor iteration t in θ_t, then $\theta_{t+1} = \theta_t + \lambda\kappa$, with λ a step-length parameter, and κ a direction vector. The restrictions on the parameters, both ultrametric (2) and external (3), are formulated as R(θ) ≥ 0. The direction κ is found by minimizing:

$$\frac{1}{2}\boldsymbol{\kappa}' \boldsymbol{H}(\boldsymbol{\theta}_t)\boldsymbol{\kappa} + \boldsymbol{g}(\boldsymbol{\theta}_t)\boldsymbol{\kappa} \tag{5}$$

subject to:

$$r(\boldsymbol{\theta}_t)\boldsymbol{\kappa} + \boldsymbol{R}(\boldsymbol{\theta}_t)\boldsymbol{\kappa} \geq 0 \ , \tag{6}$$

where H(θ) is the Hessian and g(θ) is the gradient of the likelihood, and r(θ) is the Jacobean of R(θ) with respect to θ. We approximate these derivatives numerically using forward differences. The E-M algorithm is started using unconstrained estimation in the M-step. After convergence, the ultrametric constraints in (2) are approximated for each class by the triple reduction method, which involves a repeated sequential averaging of the largest two pairs of each triple (Roux, 1987). The additional restrictions in (3) are similarly enforced by repeated averaging[1].Then, from the starting values thus obtained, the E-M cycle is repeated applying the SQP constrained estimation in the M-step, using the Broyden, Fletcher, Goldfarb and Shanno (Scales, 1985) Quasi Newton method. We use the SQP algorithm implemented in GAUSS (Aptech, 1995). The E-step of the E-M algorithm involves taking the expectation of the complete log-likelihood with respect to the unobserved class membership indicators, which amounts to replacing these indicators with their expected values. These expected values equal the posterior probabilities, π_{ns}, that subject n belongs to class s calculated at the current parameter estimatesby means of Bayes' Theorem:

$$\pi_{ns} = \frac{\pi_s \phi_s (\boldsymbol{D}_n | \boldsymbol{\Delta}_s, \boldsymbol{\Sigma}_s)}{\sum_{s=1}^{S} \pi_s \phi_s (\boldsymbol{D}_n | \boldsymbol{\Delta}_s, \boldsymbol{\Sigma}_s)} \ . \tag{7}$$

These π_{ns} provide a probabilistic allocation of the objects to the classes. For further details on the E-M algorithm we refer to Dempster, Laird, and Rubin (1977). Since the EM algorithm may converge to local optima, the algorithm needs to be started from several random starts of the posterior probabilities to minimize the probability of convergence to local optima, or some rational start, derived for example from a K-means clustering of the observed distances, may be employed.

[1] We have found the sequential averaging procedure to approximately satisfy the constraints to greatly enhance the performance of the SQP algorithm in the M-step. This arises because the averaging procedure brings the initial estimates in regions where the derivatives of the ultrametric constraints can be calculated.

2.2 Model selection and number of parameters

Since the number of classes is unknown, information type criteria, such as AIC and CAIC, can be used as heuristics to determine the number of classes (Bozdogan, 1987). These information theoretic measures impose a penalty on (minus two times) the likelihood, amounting to respectively M and M(ln(N)+1), with M the number of parameters estimated and N=np the number of observations. We investigate the separation of the classes using an entropy measure (cf. Wedel & DeSarbo, 1994). Without external constraints, the effective number of parameters estimated M=S(I-1)+2*S-1 for a model with a diagonal covariance matrix, since there are I-1 parameters for the full ultrametric tree in each class, plus S variances and S-1 priors to be estimated. For example, for the equicorrelation model S correlations are added to the number of parameters estimated. Each external constraint for a particular class in (3) partitions the class level tree of I stimuli in two subtrees of I_1 and I_2 stimuli and therefore reduces the number of tree-distance parameters estimated in that class from (I-1) to (I_1 -1+I_2-1). Since constrained and full models are nested, likelihood ratio tests may be employed to test the constraints. In the next section we investigate the performance of the algorithm to recover true ultrametric distances, as well as the performance of the AIC and CAIC statistics to identify the true number of classes.

3. Synthetic data analysis

In order to demonstrate the performance of the algorithm, we generated synthetic data with different numbers of classes and stimuli. A different set of distances, satisfying the ultrametric inequality was generated for each of the -two or four- classes in the data. These were taken from subsets of the stimuli in the Table 5.3 in DeSarbo, Manrai, and Manrai (1993). The first data set was generated for S=2, n=20 and I=5 stimuli, labelled A through E. To each class, 10 subjects were assigned. For Class 1 random error drawn from N(0, 0.1) was added to these true distances, for Class 2 the error was drawn from N(0, 0.5). The number of ultrametric constraints is 10 for each class. The second data set was generated with I=8 stimuli, labelled A through H, and S=2 classes, 10 subjects per class. Random normal error drawn from N(0,0.1) and N(0,0.5) was added to these distances in Classes 1 and 2 respectively. The number of constraints is 56 per class. The third synthetic data set pertained to I=5 stimuli, and S=4 classes. Random error drawn from respectively N(0,0.1), N(0,0.2), N(0,0.3) and N(0,0.4) was added to the distances in Classes 1 through 4. The number of constraints is 10 for each class. The fourth synthetic data set analysed is identical to data set 1 (S=2, I=5), except that the errors were drawn from N(0,1) and N(0,2, respectively. The above mixture of trees was brought to each of these four data sets from S=1 to S=5, with a diagonal covariance matrix, and without external constraints. Each analysis was repeated several times in order to detect possible local optima of the log-likelihood.

Table 1 provides the number of parameters estimated, the log-likelihood, AIC and CAIC statistics for the S=1 to S=5 solutions for these four data sets. The table shows that the AIC indicates the correct number of classes in all cases except for data set 4, CAIC indicates the correct number of classes for all four data sets. CAIC appears to be somewhat more conservative, and to be preferable to identify the appropriate number of classes (Bozdogan, 1987). Local optima were more often encountered when the number of classes specified in the analysis was larger than the true number of classes.

[INSERT TABLE 1 HERE]

Tables 2 to 5 provide the true and estimated distances, and the estimated standard errors per class for each of the four data sets. In addition, the Table shows the R^2 between true and estimated distances, and the Root-Mean-Squared-Error (RMSE) as a percentage of the mean of the true distances. The algorithm appears to perform well in recovering the parameters of the model in these synthetic data examples, as can be seen from the true and estimated distances in Tables 2 to 5. The EM algorithm converged within 7 iterations for all analyses. The Entropy of the posteriors equalled 1.0000 for all four analyses, indicating perfect separation of the classes: all subjects were assigned to their true class with posterior probability 1.0000. When the results in Table 5 are compared to those in Table 2 it may be seen that parameter recovery deteriorates somewhat as the random error in the distances increases. Comparing Tables 2 and 3 one observes that the performance of the algorithm does not seem to be affected by larger numbers of distances to be estimated and larger numbers of constraints imposed. The correlations of true and estimated distances are very close to one for all three applications, and the RMSE's are below 10% of the mean of the distances. The algorithm converged to the same solution from several random starts for the true number of classes, indicating that there were no problems of local optima. Figures 1 to 3 shows the class level trees resulting for the analyses for data sets 1 to 3 (the results for data set 4 are not shown since they are the same as those for data set 1).

[INSERT TABLES 2 TO 5 AND FIGURES 1 TO 3 HERE]

3. Application to Sensory Perception of Colas

We apply the mixture ultrametric tree model to data provided by Schiffman, Reynolds, and Young (1981, pp. 33-34). In their sensory experiment, ten different brands of cola were used: 1. Diet Pepsi (DP), 2. RC Cola (RCC), 3. Yukon (Y), 4. Dr. Pepper (DP), 5. Shasta (S), 6. Coca Cola (CC), 7. Diet Dr. Pepper (DPP), 8. Tab (T), 9. Pepsi-Cola (PC), 10. Diet Rite (DR). The colas were bought in glass containers from retail outlets. These colas were presented in 5-ounce plastic cups to 10 subjects (nonsmokers, aged 18-21 years), at room temperature, having been opened two hours before to remove carbon dioxide. Subjects were not allowed to swallow the colas and rinsed their mouths with distilled water between tastes. They were blind-folded during the experiment, and brand names were not provided to them. 45 Pairwise judgements were made with an interval of 5 minutes between pairs, the order of the pairs was randomized to balance cross-adaptation effects. The similarity judgements were provided on a graphical anchored line-scale, and transcribed on a scale from 0-100 representing same -towards 0, and different -towards 100. For a more detailed description of the experiment we refer to Schifmann, Reynolds, and Young (1981).

3.1 Results of the mixture of ultrametric trees model

Our purpose in this particular application is to investigate the existence of perceptual heterogeneity among subjects, using the proposed mixture of ultrametric trees model. We tested for S=1, 2, or 3 classes. We first estimate the independence model with a diagonal correlation matrix and investigate several alternative models later on. Table 6 shows the log-likelihood, and AIC and CAIC statistics for the S=1 to S=3 models. As can be seen from that Table, both AIC and CAIC are minimum for S=2. Figure 5 therefore presents the estimated ultrametric trees for the S=2 independence model. For comparison, the S=1 solution is presented in Figure 4. The S=1 ultrametric tree structure for the ten cola brands in Figure 4 presents a somewhat mixed picture. One branch of the tree contains three diet colas, which have relatively low path-length distances (the distance to their least common ancestor node): Diet Pepsi, Diet Rite and Tab. Under the feature matching model the path-length from the root of the tree to the least common ancestor node of these tree diet cola's is a measure of the importance of the features shared by these stimuli (cf. Corter, 1996). However, the interpretation of the common ancestor node of these three colas as diet/non-diet feature is hampered by the fact that the fourth diet cola in the stimulus set, Diet Dr Pepper is joined to the regular Dr Pepper, albeit with a relatively high path length distance. The common ancestor node of these two brands should be interpreted as a brand taste feature: Dr Pepper/other, which can be interpreted as the characteristic cherry-type flavour of Dr. Pepper colas (Schiffman, Reynolds & Young, 1981). In a similar way, the subtree in the middle of the ultrametric tree in Figure 4 shows a set of nodes that can be interpreted as representing brand-specific features, distinguishing the five remaining nondiet brands.

In the S=2 solution in Figure 5 the hierarchical structure differs substantially among the two classes. First, Class 1 clearly perceives taste differences between diet and non-diet Cola's. Diet Pepsi and Diet Rite have a very low path length distance, indicating very similar tastes. These two brands form one subtree, together with Diet Dr. Pepper and the diet version of Coca Cola, at the time of the study, Tab. The common ancestor node of these four brands can therefore be interpreted as diet/nondiet taste feature. The path length from the root to the common ancestor node indicates that this is the most important feature determining similarity judgements of the subjects in this class. Further, observe that Diet Dr. Pepper shares the least number of features with the other brands in this subtree. The corresponding node can be interpreted as cherry/non-cherry flavour. In the non-diet subtree various colas are joined at different path-lengths, but here too Dr Pepper seems to stand out (together with RC cola). The second Class appears to separate brands in several subtrees: Dr. Pepper and Diet Dr Pepper, respectively Coca Cola and its diet version Tab, are joined at relatively low path length distances, indicating a large number of common features amongst them. This class of subjects seems to primarily taste differences among brands. Note that particularly the Dr. Pepper brands stand out, the length of the path from the root to the node indicating that the cherry/non cherry flavour feature is the most important distinguishing characteristic for subjects in this class. However, the exception to this grouping of brands is that Diet Pepsi and Regular Pepsi are joined in a subtree with several other brands. Nevertheless, we conclude that the specific brand tastes are the dominant features determining similarity judgements in this class. Comparing the S=2 results with the S=1 of Figure 4, it is obvious that the heterogeneity in the sample masks the ultrametric tree structure at the aggregate level and causes the mixed structure in the S=1 solution, rendering it less interpretable. In order to further investigate the fit of the S=2 unrestricted independence mixture of trees, an analysis of the residuals was conducted that showed that 96.4% of the standardised residuals was in between -2 and 2, and all residuals were below 2.37 in absolute value. The curtosis of the residuals is 3.05 (se=0.23), and the skewness is -1.30 (se=0.12). Thus, the residuals appear to be somewhat skewed to the left and to have a somewhat higher curtosis as compared to the normal distribution, but there were no indications of outliers[2] .

The two classes in Figure 5 are very well separated and the posterior probabilities equal zero or one up to five decimals, the entropy criterion E=1.0000. Class 1 consists of subjects 1, 4, 5, 6 and 9, and Class 2 of the other five subjects. Interestingly, all of the subjects in Class 1 have the ability to taste a bitter tasting compound called PTC, while the subjects in Class 2 (2, 3, 7, 8, and 10) do not have that ability (Schiffman, Reynolds & Young, 1981). This physiological characteristic discriminates perfectly between the two classes, and apparently subjects ability to taste PTC determines the extent to which they use diet/ non-diet, or specific brand tastes as the dominant feature to determine similarities.

3.2 Alternative model tests

In order to further investigate the feature matching tree structures of the two classes, and to illustrate the use of external constraints, we impose additional restrictions upon the ultrametric distances in each of the classes, to test the above hypothesis of a diet/nondiet first versus a brand first structure in the two segments. In particular, for Class 1 we constrain the

[2] In order to investigate whether a log-normal distribution provides a better fit to these data, the same S=2 model was estimated to the data after taking the natural logarithm of the observed distances (and adding one to prevent taking the log of zero). Skewness and curtosis were 3.15 and -1.29, respectively, while 1.6% of the residuals was below -3, none were above 3. The minium residual value was -4.201. Clearly, the log-normal distribution does not provide a better fit to these data.

non-diet colas and diet colas to be in distinct subtrees by restricting the distances from each of the diet cola's (brands 1, 7, 8, and 10) to a particular non-diet cola (brands 2, 3, 4, 5, 6, 9) to be equal. For Class 2 we constrain the Pepsi cola's (1 and 9), the Coca cola's (6 and 8) and the Dr Pepper cola's (4 and 7) to be in distinct subtrees, by constraining the distances from the two brands in each of these pairs to all other brands to be equal (see equation 3). Note that most of the constraints were already satisfied in the unrestricted solution, except for the restriction that forces the Pepsi cola's in the same branch of the tree in Class 2. Therefore, in this constrained model there are 9 free parameters in Class 1, and in Class 2 there are 8 free parameters, bringing the total number of parameters in the restricted model to 21. For the restricted model, the LR test statistic relative to the unrestricted model above is 4.8837 with 1 df. This test is significant (AIC and CAIC also favour the unrestricted model, see Table 6). Figure 6 presents the restricted ultrametric trees in the two classes. It shows that the ultrametric tree for Class 1 is the same as that for the unrestricted solution presented in Figure 5, but that the ultrametric tree in Class 2 is changed relative to the unrestricted solution, because of the constraint that Pepsi and Diet Pepsi should be in the same subtree. In particular, Diet Pepsi, and Pepsi, respectively Diet Rite and Shasta are joined in two subtrees, and the model thus forces one common feature for the brands in these subtrees. Apparently, this is an oversimplification of the feature structure that subjects in this class use to determine similarity between colas.

Next, we investigate whether a model that allows for correlations among the dissimilarities provides a better fit. Since the order in which the dissimilarities were obtained is unknown for these published data, the AR(1) model presented in Section 2 above cannot be specified, but the equicorrelation model presents a parsimonious and plausible alternative for the covariance structure among the dissimilarity judgements. The equicorrelation model states that dissimilarity judgements provided by the same individual, given class s, exhibit a constant correlation of ρ_s, while being uncorrelated among subjects. Such a model is similar in spirit to hierarchical models where the variances of measurements are assumed to differ within and between subjects. The LR test for the S=2 equicorrelation model against the S=2 independence model has a Chi-squared value of 7.270, with 2 df. (P=0.013), indicating that the equicorrelation model fits somewhat better than the independence model (in addition AIC is lower for the equicorrelation model and CAIC is slightly lower). The estimated correlations in the equicorrelation model were 0.025 for Class 1 and 0.089

for Class 2, the estimated standard errors were σ_1=22.983 and σ_2=21.654 respectively. Both estimated correlations are quite low. Nevertheless, there seems to be some indication of a correlation between the dissimilarities provided by the same individual as indicated by the likelihood ratio test. The fact that the correlations are low can probably be attributed to the design of the experiment, in which orders were randomized and subjects rinsed their mouths with distilled water after each judgement. The estimated ultrametric distances of the S=2 equicorrelation model are very close to those of the independence model (the differences are in the decimals), and yield the same ultrametric trees as presented for the independence model in Figure 5. The correlations of observed and estimated distances are 0.827 for Class 1, and 0.903 for Class 2. Figure 6 provides a scatterplot of observed and estimated distances. The posterior probabilities are unchanged, and the residual analysis yields very similar results to that of the independence model presented above. We conclude that the unrestricted S=2 equicorrelation model is a reasonable approximation to the data. Table 7 presents the estimated distances of this model and the averages of the observed distances in each of the two classes.

4. Conclusions

The analyses of synthetic and empirical data aptly demonstrate the performance of our proposed procedure to estimate the mixture of trees. The E-M algorithm converged to its final solution in a rather small number of steps (below 20 in most cases). The estimation algorithm, in the applications presented, seemed not to suffer from local optima, at least when the number of classes estimated equalled the true number of classes. The fast convergence of the E-M algorithm and the fact that there were no serious problems of local optima are probably caused by the fact that there is a large number of paired comparison observations for each subject in such applications, thus providing much information on the posterior update in each E-step. The likelihood information dominates the prior information in the Bayesian posterior calculations, due to which local optima are probably less likely to occur and the algorithm converges quickly. We have found that approximately enforcing the constraints on the initial estimates by repeated averaging is essential for good performance of the SQP method in the M-step, especially when such a large number of constraints is enforced as in ultrametric tree models.

The results revealed by our restricted mixture of ultrametric trees methodology in its application to the Schiffman, Reynolds and Young (1981) data are interesting and encouraging. Classes with a different hierarchical structure of common features, evidenced by different tree-topologies, were identified. The two ultrametric trees arose from the difference in the relative importance of diet/non-diet versus brand-specific taste-features. The test for the restricted model enforcing such brand-first versus diet/non-diet first topologies on the trees for the two classes showed that this one brand (Pepsi) may be an exception, and has been able to differentiate its diet version with a specific taste feature. In addition, membership to the diet/non-diet first class is entirely attributable to the ability of subjects to taste the bitter chemical compound PTC. We note that although the number of subjects in the application was rather small, the total number of observations is large, due to the large number of paired comparisons.

Our approach to deal with heterogeneity in ultrametric tree representations of subjects decision process seems to provide valuable insights into the formation of similarity judgements by subjects, and differences among them. In addition, the possibility to impose constraints on the tree topology allows one to test various hypotheses about that topology, and allows subjects to be assigned to a class of which the hypothesized tree structure has the highest posterior probability of having generated their paired comparison evaluations. In the application to the Schiffman, Reynolds and Young (1981) data, the posterior probabilities for the mixtures of ultrametric trees with and without external constraints were identical, but this obviously need not be the case in all applications..

References

Allenby, G.M. (1989). A unified approach to identifying, estimating and testing demand structures with aggregate scanner data. *Marketing Science*, 8, 265-280.

Aptech (1995). *Constrained maximum likelihood, GAUSS manual*. Aptech systems, Maple Valley.

Bozdogan, H. (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345-370.

Corter, J.E. (1996). *Tree models of similarity*. London: Sage.

Corter J.E., & Tversky, A. (1986). Extended Similarity Trees. *Psychometrika* 51, 429-451.

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

Carroll J.D. & Pruzanski, S. (1980). Discrete and hybrid scaling models. In E.D. Lanterman and H. Feger (eds), *Similarity and Choice*, Hans Huber, Bern, pp. 48-69.

Carroll, J.D., Clark, L. & DeSarbo, W.S. (1984). The representation of three-way proximities data by single and multiple tree structure models. *Journal of Classification*, 1, 25-74.

Dempster, A. P., Laird, N. M. & Rubin, R. B. (1977) Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society*. B39, 1-38.

DeSarbo, W.S. & Mahajan, V. (1984). Constrained classification: the use of a priori information in cluster analysis. *Psychometrika*, 49, 187-215.

DeSarbo, W.S., Manrai, A.K. & Manrai, L.A. (1993). Non-Spatial Tree Models for the Assessment of Competitive Market Structure: An Integrated Review of the Marketing and Psychometric Literature. In: J Eliashberg & G.L. Lilien (eds). *Marketing, Handbooks of OR&MS*, 5, 193-257.

DeSoete, G. (1984). A least squares algorithm for fiiting trees to proximity data. *Psychometrika,* 48, 621-626.

Ferligoj, A. & Batagelj, V. (1982). Some types of clustering with relational constraints. *Psychometrika*, 47, 541- 552.

Gordon, A.D. (1973). Classification in the presence of constraints. *Biometrics*, 29, 821-827.

Gordon, A.D. (1980). *Classification.* Chapman and Hall, London.

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

Jamshidian, M. & Bentler, P.M. (1993). A modified Newtone method for constrained estimation in covariance structure analysis. *Computational Statistics and Data Analysis*, 15, 133-146.

Johnson, S.C. (1967). Hierarchical Clustering Schemes. *Psychometrika*, 32, 241-254.

Morgan, B.T, & Ray, A.P.G. (1995). Non-uniqueness and inversion in cluster analysis. *Applied Statistics*, 44, 117-134.

Rao, V.R. & Sabavala, D.J. (1981). Inference of hierarchical choice processes from panel data. *Journal of Consumer Research*, 8, 85-96.

Roux, M. (1987). Techniques of Approximation for Building Two Tree Structures. *Proc. Franco-Japanese Scientific Seminar: Recent Developments in Clustering and Data-Analysis*, Tokyo, pp. 127-146.

Sattah S., & Tversky, A. (1977). Additive Similarity Trees. *Psychometrika* 42, 319-345.

Scales, L.E. (1985). *Introduction to non-linear optimization.* MacMillan, London.

Schiffman, S.S., Reynolds, M.L. & Young, F.W. (1981). Introduction to multidimensional Scaling. London: Academic Press.

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

Wedel, M., & DeSarbo, W.S. (1994), A Review of Recent Developments in Latent Class Regression Models. In: R.P. Bagozzi (ed.), *Advanced Methods of Marketing Research,*, Cambridge: Blackwell, 352-388.

Wedel, M., & DeSarbo, W.S. (1996). An exponential family mixture MDS methodology for simultaneous segmentation and product positioning. *Journal of Business and Economic Statistics, forthcoming.*

Wedel, M., & DeSarbo, W.S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12, 21-56.

Table 1. Fit Statistics for the S=1 to S=5 models for the synthetic data

	S=1	**S=2**	**S=3**	**S=4**	**S=5**
Dataset 1 (I=5,S=2)					
No. Parameters	5	11	17	23	29
log-l	-561.4278	12.8623	12.8623	12.8625	12.8623
AIC	1127.8557	-14.7247*	-8.7247	-2.7252	3.2754
CAIC	1153.3472	41.5568*	78.3467	115.13607	151.9266
Dataset 2 (I=8,S=2)					
No. Parameters	8	17	26	35	44
Log-l	-1814.0094	39.8791	39.8791	39.8791	39.8791
AIC	3636.0189	-62.7582*	-53.7582	-44.7582	-35.7582
CAIC	3685.6423	42.8168*	107.7682	172.7196	237.6711
Dataset 3 (I=5,S=4)					
No. Parameters	5	11	17	23	29
Log-l	-1257.3560	-933.6958	-654.2448	40.9144	40.9144
AIC	2519.7120	1878.3917	1325.4895	-58.8288*	-52.8288
CAIC	2549.6694	1944.2977	1427.3444	78.9749*	120.9236
Dataset 4 (I=5,S=2)					
No. Parameters	5	11	17	23	29
Log-l	-579.4465	-361.5404	-349.7865	-349.7930	-352.8856
AIC	1163.8930	734.0808	716.5731*	722.5860	734.7712
CAIC	1190.3846	792.3622*	806.6445	844.4473	888.4224

* Denotes the minimum value of the statistic for that dataset.

Table 2. True (lower) and estimated (upper) distances for the I=5, S=2 model

Class 1	**A**	**B**	**C**	**D**	**E**
A	-	23.462	23.462	23.462	22.371
B	23.48	-	20.808	20.808	23.462
C	23.48	20.79	-	1.595	23.462
D	23.48	20.79	1.56	-	23.462
E	22.37	23.48	23.48	23.48	-
Class 2	**A**	**B**	**C**	**D**	**E**
A	-	23.488	20.967	12.971	16.772
B	23.48	-	23.488	23.488	23.488
C	20.79	23.48	-	20.967	20.967
D	12.99	23.48	20.79	-	16.772
E	16.74	23.48	20.79	16.74	-

r^2=0.9999; σ_1=0.115, σ_2 =0.448 RMSE=0.3464%

Table 3. True (Lower) and Estimated (Upper) Distances for the I=8, S=2 Model

Class 1	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	-	23.469	23.469	23.469	22.431	23.469	23.469	23.469
B	23.48	-	20.808	20.808	23.469	18.034	20.808	20.808
C	23.48	20.79	-	1.536	23.469	20.808	13.029	16.748
D	23.48	20.79	1.56	-	23.469	20.808	13.029	16.748
E	22.37	23.48	23.48	23.48	-	23.469	23.469	23.469
F	23.38	18.07	20.79	20.79	23.48	-	20.808	20.808
G	23.48	20.79	12.99	12.99	23.48	20.79	-	16.748
H	23.48	20.79	16.74	16.74	23.48	20.79	16.74	-
Class 2	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	-	16.726	20.739	23.504	16.726	16.726	20.739	23.504
B	16.74	-	20.739	23.504	12.975	12.975	20.739	23.504
C	20.79	20.79	-	23.504	20.739	20.739	18.190	23.504
D	23.48	23.48	23.48	-	23.504	23.504	23.504	22.461
E	16.74	12.99	20.79	23.48	-	1.580	20.739	23.504
F	16.74	12.99	20.79	23.48	1.56	-	20.739	23.504
G	20.79	20.79	18.07	23.48	20.79	20.79	-	23.504
H	23.48	23.48	23.38	22.37	23.48	23.48	23.48	-

r^2=1.0000; σ_1 = 0.106; σ_2 =0.481; RMSE=0.1952%

Table 4. True (Lower) and Estimated (Upper) Distances for the I=5. S=4 Model

Class 1	**A**	**B**	**C**	**D**	**E**
A	-	23.466	23.466	23.466	22.437
B	23.48	-	20.815	20.815	23.466
C	23.48	20.79	-	1.584	23.466
D	23.48	20.79	1.56	-	23.466
E	22.37	23.48	23.48	23.48	-
Class 2	**A**	**B**	**C**	**D**	**E**
A	-	20.747	20.747	23.463	18.122
B	20.79	-	1.530	23.463	20.747
C	20.79	1.56	-	23.463	20.747
D	23.48	23.48	23.48	-	23.463
E	18.07	20.79	20.79	23.48	-
Class 3	**A**	**B**	**C**	**D**	**E**
A	-	1.604	23.426	20.808	12.952
B	1.56	-	23.426	20.808	12.952
C	23.48	23.48	-	23.426	23.426
D	20.79	20.79	23.48	-	20.808
E	22.99	12.99	23.48	20.79	-
Class 4	**A**	**B**	**C**	**D**	**E**
A	-	23.508	20.762	12.980	16.788
B	23.48	-	23.508	23.508	23.508
C	20.79	23.48	-	20.762	20.762
D	12.99	23.48	20.79	-	16.772
E	16.74	23.48	20.79	16.74	-

r^2=0.9650; σ_1=0.104; σ_2 =0.206;σ_3 =0.276;σ_4 =0.386;RMSE=7.9405%.

Table 5. True (Lower) and Estimated (Upper) Distances for the I=5. S=2 Model (high error)

Class 1	**A**	**B**	**C**	**D**	**E**
A	-	23.283	23.283	23.283	22.154
B	23.48	-	20.328	20.328	23.283
C	23.48	20.79	-	2.178	23.283
D	23.48	20.79	1.56	-	23.283
E	22.37	23.48	23.48	23.48	-
Class 2	**A**	**B**	**C**	**D**	**E**
A	-	22.902	21.534	13.769	15.563
B	23.48	-	22.902	22.902	22.902
C	20.79	23.48	-	21.534	21.534
D	12.99	23.48	20.79	-	15.563
E	16.74	23.48	20.79	16.74	-

r^2=0.9938; σ_1=0.951;σ_2=2.289; RMSE=2.9859%

Table 6. Fit statistics for several mixture tree models for the Schiffman et al (1981) cola data

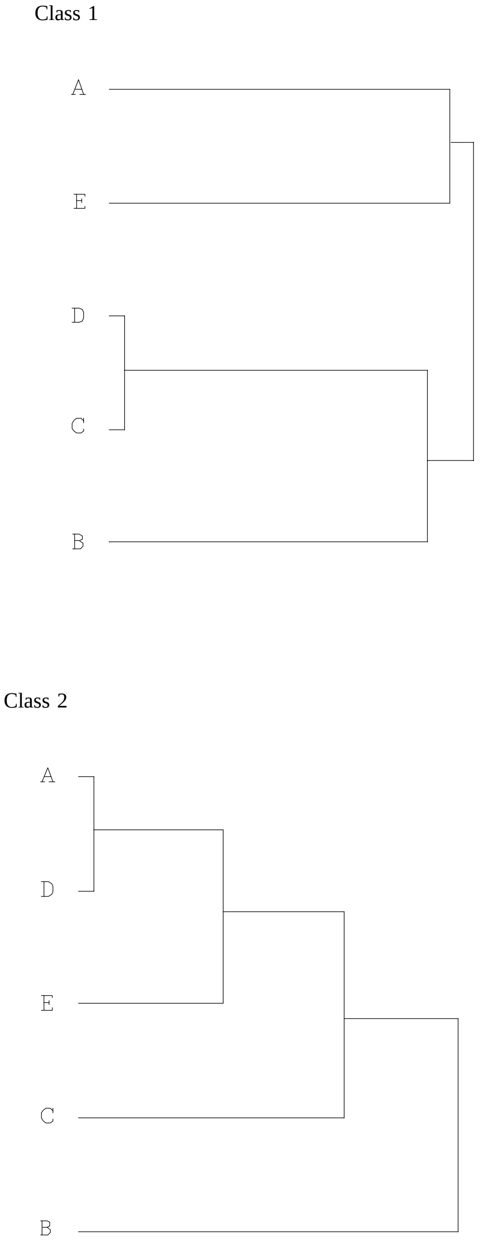
Model	df	log-l	AIC	CAIC
S=1. Unrestricted. Independence	10	-2098.8203	4207.6406	4268.7331
S=2. Unrestricted. Independence	21	-2035.7535	4092.5070*	4220.8012*
S=3. Unrestricted. Independence	32	-2034.2425	5000.4850	4295.9810
S=2. Restricted. Independence	20	-2040.6372	4101.2744	4223.4594
S=2. Unrestricted. Equicorrelation	23	-2028.4840	4079.9679#	4220.4801#

* denotes minimum value across S=1. 2. 3; # denotes minimum value across all models

Table 7. Observed (Lower) and Estimated (Upper) Distances for the Schiffman (1981) data

Class 1	**DP**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DP	-	76.096	76.096	76.1	76.096	76.096	41.466	30.344	76.096	13.790
2	39.40	-	48.775	34.51	48.775	48.775	76.096	76.096	48.775	76.096
3	85.8	39.6	-	48.78	44.498	44.498	76.096	76.096	44.498	76.096
4	84.4	24.6	70.0	-	48.775	48.775	76.096	76.096	48.775	76.096
5	79.4	30.8	42.2	46.8	-	27.268	76.096	76.096	22.206	76.096
6	67.2	33.4	26.0	91.6	28.2	-	76.096	76.096	27.268	76.096
7	30.8	85.6	65.2	79.4	67.6	88.0	-	41.466	76.096	41.466
8	62.6	89.6	72.2	83.8	77.0	66.8	39.2	-	76.096	30.344
9	75.0	22.6	51.8	44.4	22.2	23.4	66.2	85.0	-	76.096
10	13.0	79.0	81.6	84.4	81.0	66.2	41.6	20.0	85.6	-
Class 2	**DR**	**RCC**	**Y**	**DRP**	**S**	**CC**	**DDP**	**T**	**PC**	**DR**
DR	-	46.660	63.949	84.479	46.660	55.650	84.479	55.650	46.660	46.660
RCC	30.2	-	63.949	84.479	28.325	55.650	84.479	55.650	28.325	39.461
Y	72.6	69.2	-	84.479	63.949	63.949	84.479	63.949	63.949	63.949
DRP	87.6	87.4	71.0	-	84.479	84.479	20.795	84.479	84.479	84.479
S	73.2	30.2	60.2	85.8	-	55.650	84.479	55.650	22.996	39.461
CC	59.4	48.0	49.6	88.4	42.6	-	84.479	41.395	55.650	55.650
DDP	85.0	86.4	90.2	20.8	84.4	66.2	-	84.479	84.479	84.479
T	60.6	71.8	71.0	93.4	58.0	41.4	93.0	-	55.650	55.650
PC	56.2	24.6	87.0	88.0	23.0	46.8	87.4	57.6	-	39.461
DR	39.0	42.8	58.4	94.4	45.2	69.6	77.0	47.2	33.0	-

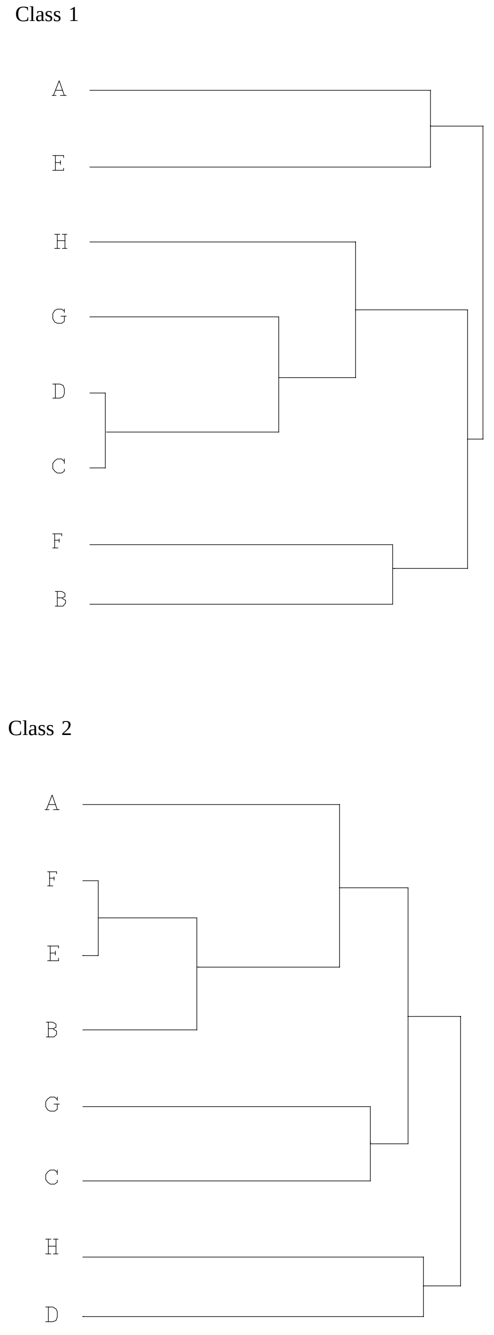
Figure 1. Recovered Trees for S=2. I=5 Synthetic data analysis.



Class 1

Class 2

Figure 2. Recovered Trees for S=2. I=8 Synthetic Data Analysis.



Class 1

Class 2

Figure 3. Recovered Trees for S=4 I=5 Synthetic Data Analysis.

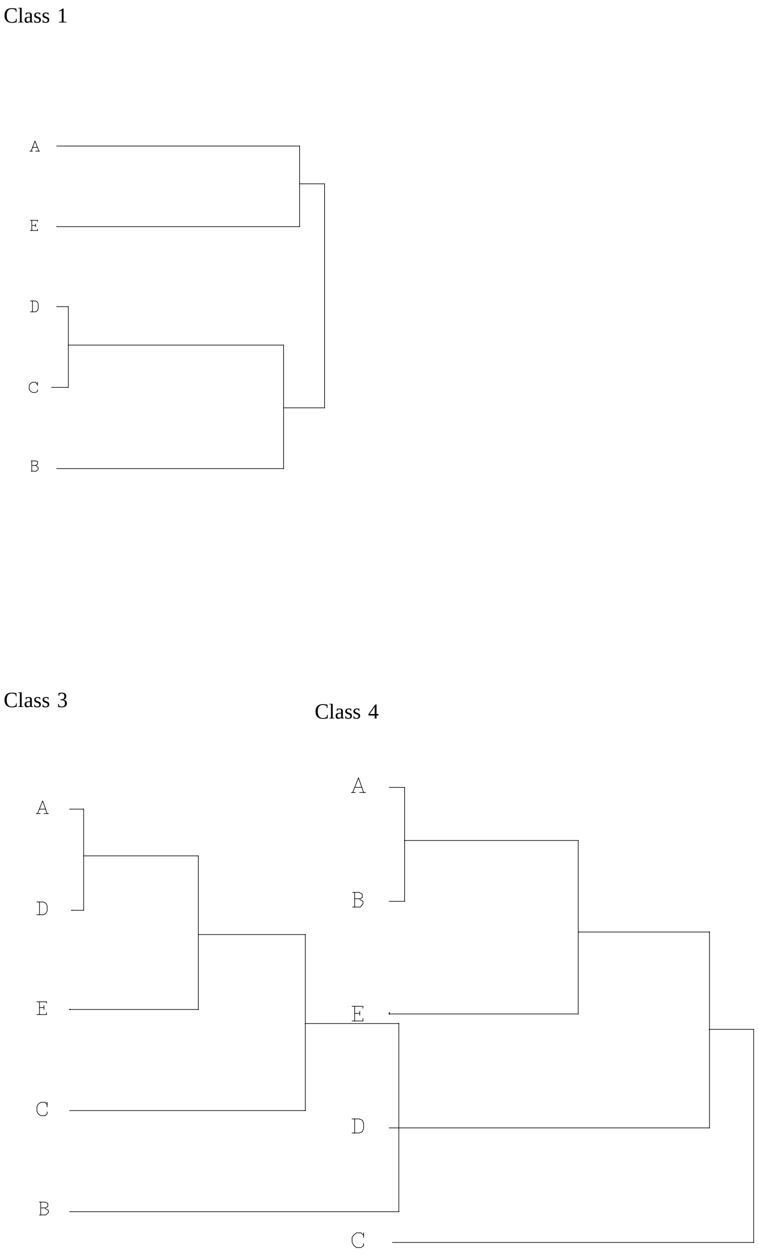


Figure 4. Ultrametric tree (S=1. independence) Schiffman et al (1981) cola data

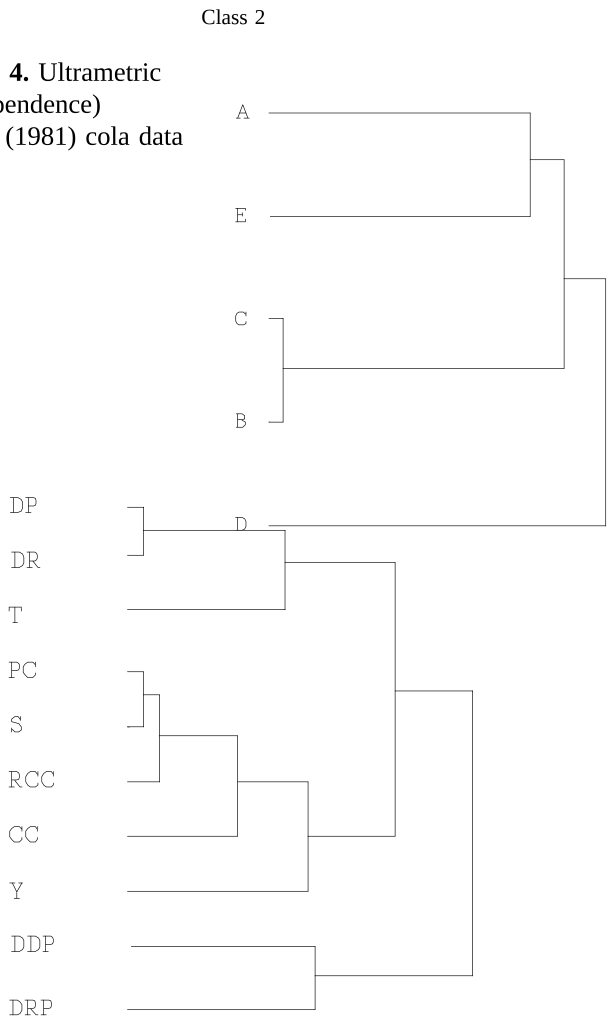


Figure 5. Ultrametric trees (S=2. independence) Schiffman et al (1981) cola data.

Class 2

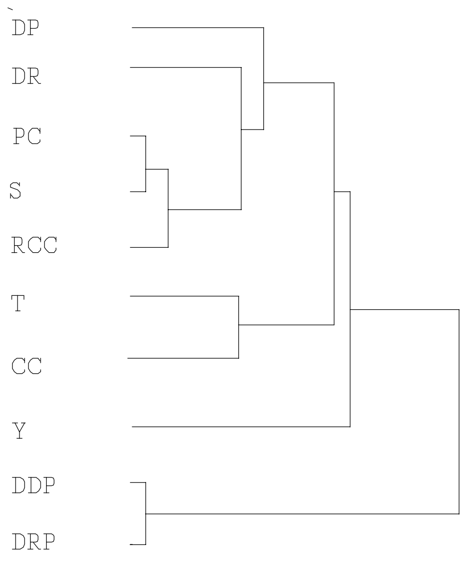


Figure 6. Constrained ultrametric trees (S=2. independence) Schiffman et al (1981) cola data.

Class 1

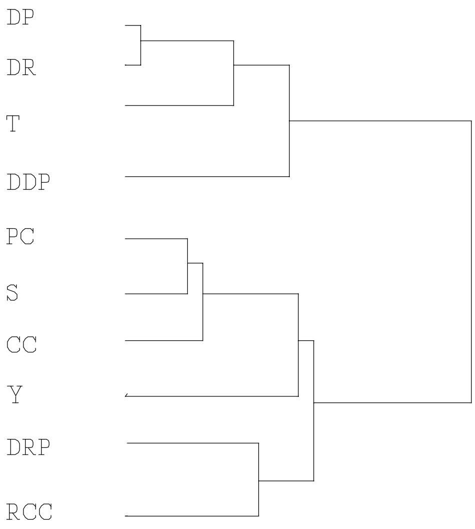
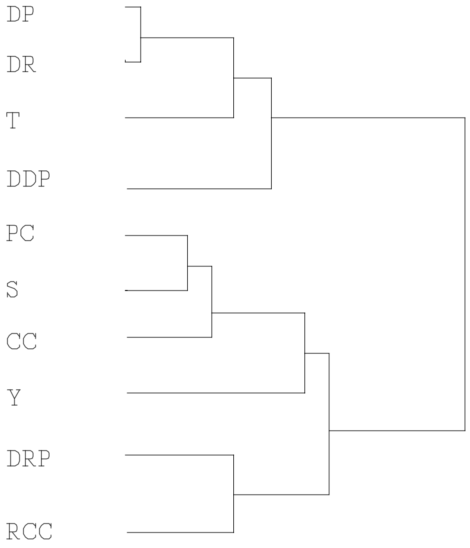
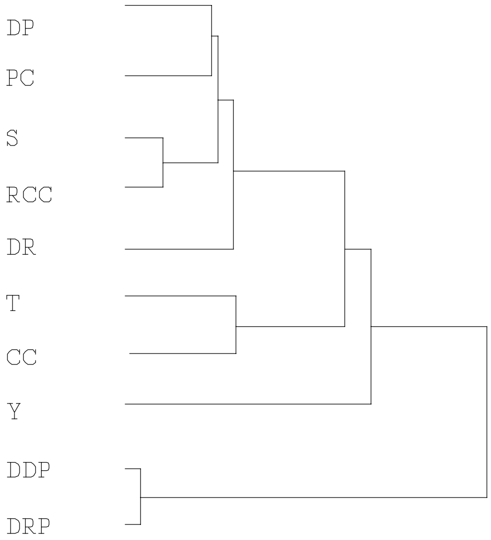


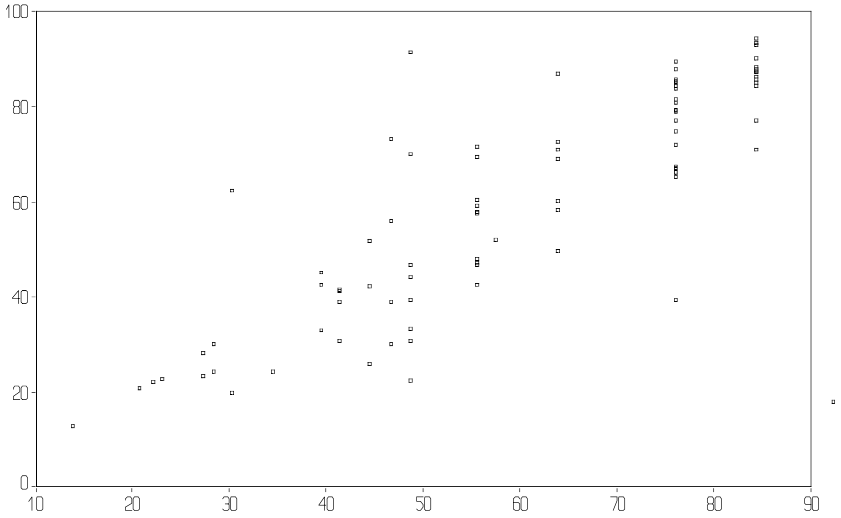
Figure 7. Observed (Y-axis) versus Estimated (X-axis) Distances

Class 1



Class 2




100
80
60
40
20
0
10
20
30
40
50
60
70
80
90