4TH INTERNATIONAL WORKSHOP ON COGNITIVE INFORMATION PROCESSING, MAY 26–28, 2014, COPENHAGEN, DENMARK

DISCOVERING HIERARCHICAL STRUCTURE IN NORMAL RELATIONAL DATA

Mikkel N. Schmidt, Tue Herlau, and Morten Mørup

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Kgs. Lyngby, Denmark

ABSTRACT

Index Terms— Hierarchical clustering is a widely used tool for structuring and visualizing complex data using similarity. Traditionally, hierarchical clustering is based on local heuristics that do not explicitly provide assessment of the *statistical saliency* of the extracted hierarchy. We propose a non-parametric generative model for hierarchical clustering of similarity based on multifurcating Gibbs fragmentation trees. This allows us to infer and display the posterior distribution of hierarchical structures that comply with the data. We demonstrate the utility of our method on synthetic data and data of functional brain connectivity.

1. INTRODUCTION

Clustering is the task of organizing a set of data objects into groups, such that objects within a group are similar in some sense. This unsupervised learning method is often used for explorative data analysis. There is not one single approach to clustering that is best in all circumstances: How objects should be clustered depends on the objective of the cluster analysis. Thus, when performing cluster analysis it is of great importance to choose appropriate clustering criteria, i.e. an algorithm, a similarity measure, or a statistical model.

Some of the most popular clustering methods are k-means clustering, hierarchical clustering, and mixture models. In k-means clustering a centroid is computed for each cluster by computing the mean of all data objects belonging to the cluster. Data objects are then reassigned to the closest centroid, and the centroids are recomputed. This is iterated until convergence [1]. The k-means algorithm thus relies on a similarity measure and the ability to compute a mean value in the data space. A mixture model can be seen as a statistical extension of the k-means algorithm. Here, each cluster is endowed with a probability distribution instead of a centroid (for example a Gaussian distribution characterized by its mean and covariance), and the "similarity measure" is the probability of belonging to each cluster. When using a statistical model, in addition to discovering the best clustering it is also possible to infer the uncertainty related to the clustering, for example in form of a Bayesian posterior distribution over clusterings. In both k-means and mixture modeling, the number of clusters is usually specified beforehand, but extensions of both methods exist in which the number of clusters is learned from data [2].

In (agglomerative) hierarchical clustering each data object is initially in its own cluster. The two most similar clusters are merged and this process is repeated until all data objects are in the same cluster. The dissimilarity between two clusters is computed using a linkage function such as the smallest (single linkage), average (average linkage), or largest (complete linkage) dissimilarity between two data objects in the two clusters. The hierarchical clustering algorithm thus depends only on the choice of linkage function and similarity/dissimilarity measure. The output of the algorithm is a hierarchy (i.e., a binary tree), often represented graphically as a dendrogram in which the height denotes the dissimilarity value at which two clusters were merged. In one sense, the hierarchical clustering approach gives more information about the structure in the data compared with approaches which result in (a posterior distribution over) a single clustering: The dendrogram represents a large number of different clusterings compatible with the data at different scales or resolutions. On the other hand, the dendrogram does not say at which resolution the clustering is most salient. This problem can be solved by taking a statistical approach to hierarchical clustering [3]: The ad hoc similarity measure can be replaced by a suitable probability distribution, and clusters can be merged based on how well they fit together according to the statistical model. The statistical merge criteria can then be used to assess the saliency of each level of the hierarchy.

All of the methods mentioned above aim at grouping data objects together that are *directly similar* as measured by a similarity measure or as plausibly being generated from the same probability distribution. A different approach consists of grouping data objects which are *structurally similar*, in the sense that they have similar relations to all other objects. This is the idea behind the socalled *stochastic blockmodel* [4, 5]. Consider an example where we compute a correlation coefficient to compare each data object. In the former approach, data objects which are highly correlated would be grouped together whereas in the latter approach, data objects which exhibit similar correlation patterns to all other data objects would be grouped together.

In this paper we develop a probabilistic method for hierarchical cluster analysis based on structural similarity. Our method is not restricted to inferring a binary hierarchy but allows the hierarchical structure to *multifurcate*, i.e., split into an arbitrary number of groups at each level of the tree. The output of our method is thus a posterior distribution over multifurcating hierarchies. We propose a method to graphically visualize the most likely hierarchical clustering as well as its credibility. We demonstrate the method on two toy data examples as well as a data set of correlations between brain regions.

2. METHOD

The starting point for our method is a relational data set, X, for example in form of a matrix of pair-wise similarities for a number of data objects. To formulate a statistical model for a multifurcating hierarchical clustering we specify a likelihood function that determines the probability of observing the relational data for a given hierarchical structure, and a prior distribution over the hierarchy. We denote

the hierarchical structure by T. It is a multifurcating tree which has a root node, a number of internal nodes, and N leaf nodes corresponding to the data objects. The root and internal nodes all have at least two child nodes.

2.1. Likelihood

The multifurcating hierarchy determines the dependency structure of the likelihood. Consider the hierarchy given in Fig 2.A. Each internal node in the tree T, including the root, determines a division of data objects into subgroups. At the root level, for example, the 16 data objects are split into three groups, which we will denote $s_{\text{root}} = \{g_1 = \{1, \ldots, 4\}, g_2 = \{5, \ldots, 8\}, g_3 = \{9, \ldots .16\}\}$. We model the relational data between each pair of groups as independent with the following parametrization:

$$p(X|T,\theta) = \underbrace{\prod_{s\in T}}_{\text{All internal nodes}} \underbrace{\prod_{\{g_A,g_B\}\in s}}_{\text{All pairs of groups}} \underbrace{\prod_{\{i\in g_A, j\in g_B\}}}_{\text{All pairs of objects}} p(x_{i,j}|\theta_{g_A,g_B}, T), \quad (1)$$

where we condition on θ_{g_A,g_B} which is some set of parameters that govern the distribution of the relations between nodes in group g_A and g_B.

In the following, we use Normal distributions for the likelihood, but we note that it is fairly simple to substitute with another distribution if needed [6, 7, 8]. The Normal distribution has two parameters, mean and variance, i.e., $\theta_{g_A,g_B} = \{\mu_{g_A,g_B}, \sigma^2_{g_A,g_B}\}$. For these, we adopt a conjugate Normal-Gamma prior, $p(\theta_{g_A,g_B})$, with location m, precision k, shape a, and scale b. Due to the conjugacy we can analytically integrate out θ_{g_A,g_B} yielding the following expression for the likelihood terms:

$$\int \prod_{\{i\in g_A, j\in g_B\}} p(x_{i,j}|\theta_{g_A,g_B}, T) p(\theta_{g_A,g_B}) \mathrm{d}\theta_{g_A,g_B} = \frac{\Gamma(a_n)}{\Gamma(a)} \frac{b^a}{b_n^{a_n}} \sqrt{\frac{k}{k_n}} (2\pi)^{-\frac{n}{2}}, \quad (2)$$

where

$$a_n = a + \tfrac{n}{2}, \qquad b_n = b + \tfrac{1}{2} \sum_{\{i\in g_A, j\in g_B\}} (x_{i,j} - \bar{x})^2 + \frac{kn(\bar{x} - m)^2}{2(k+n)}, \quad (3)$$

$$k_n = k + n, \qquad \bar{x} = \sum_{\{i\in g_A, j\in g_B\}} x_{i,j}, \qquad n = |g_A| \cdot |g_B|. \quad (4)$$

2.2. Prior

As a prior over the hierarchical tree structure we use a multifurcating Gibbs fragmentation tree (GFT) distribution [9, 10]:

$$P(T) = \prod_{s\in T} \frac{(\alpha+\beta)\alpha^{|s|-2}}{\frac{\Gamma(N_s+\beta)}{\Gamma(1+\beta)} - \frac{\Gamma(N_s-\alpha)}{\Gamma(1-\alpha)}} \frac{\Gamma(|s| + \frac{\beta}{\alpha})}{\Gamma(2+\frac{\beta}{\alpha})} \prod_{g\in s} \frac{\Gamma(|g| - \alpha)}{\Gamma(1-\alpha)}, \quad (5)$$

where $N_s = \sum_{g\in s} |g|$. The GFT has two parameters, α and β ($0 \le \alpha < 1$ and $\beta > -2\alpha$). For the limiting case where $\alpha = 0$ the expression reduces to

$$P(T) = \prod_{s\in T} \frac{\beta^{|s|-1}}{\frac{\Gamma(N_s+\beta)}{\Gamma(1+b)} - \Gamma(N_s)} \prod_{g\in s} \Gamma(|g|). \quad (6)$$

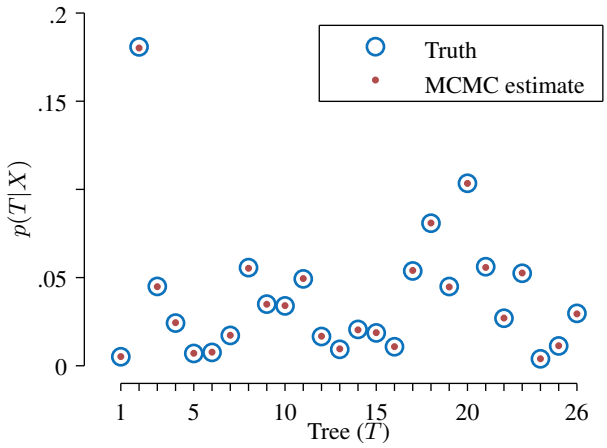


Fig. 1. Empirical evaluation of correctness of the MCMC sampler.

2.3. Inference using Markov chain Monte Carlo

It is not tractable to infer the posterior distribution of the hierarchical tree by summing over all trees, since the number of possible multifurcating trees is huge even for a relatively small number of leafs [11]. Thus, we resort to Markov chain Monte Carlo (MCMC) to infer the posterior distribution over the tree structure.

We use a Metropolis-Hastings algorithm with a subtree regrafting proposal: For a given tree a leaf node or internal node (excluding the root) is chosen uniformly at random. The subtree which has the chosen node as root is removed from the tree. If the chosen node had only a single sibling in the original tree, this sibling and its parent are combined into a single node to ensure that no internal node has less than two child nodes. The removed subtree is then inserted uniformly at random in the remaining tree. For each internal node, the subtree can either be inserted as a new child or as a new sibling. For each leaf node, the subtree can be inserted only as a new sibling (since leaf nodes, corresponding to data objects, can have no children). The probability of this proposal is found by counting the number of possible subtrees which could be removed and multiplying by the number of possible places the subtree could be inserted. The move is then accepted or rejected according to the Metropolis-Hastings acceptance ratio.

2.4. Validation of MCMC sampler

To validate the MCMC sampler we considered a simple data example with 4 data objects with the following relation matrix

$$X = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 4 & 5 \\ 2 & 4 & 0 & 6 \\ 3 & 5 & 6 & 0 \end{bmatrix}. \quad (7)$$

With only 4 data objects, the total number of possible multifurcating trees is 26 (see [11]). We computed the exact posterior distribution over these 26 solutions and compared the result with the output of the MCMC sampler, where we averaged over 1 million posterior samples. The result in Fig. 1 shows that the MCMC approximation is very precise, indicating that the tree regrafting sampler is correct.

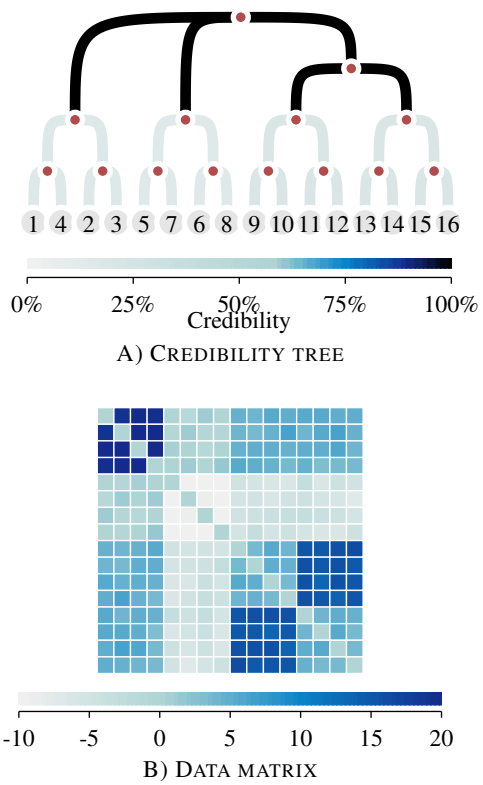


Fig. 2. Analysis of toy example 1 in which the data is split into three groups, one of which is further split into two subgroups. The output of the proposed cluster analysis is a hierarchical clustering where only the structure present in the data is assigned a high credibility.

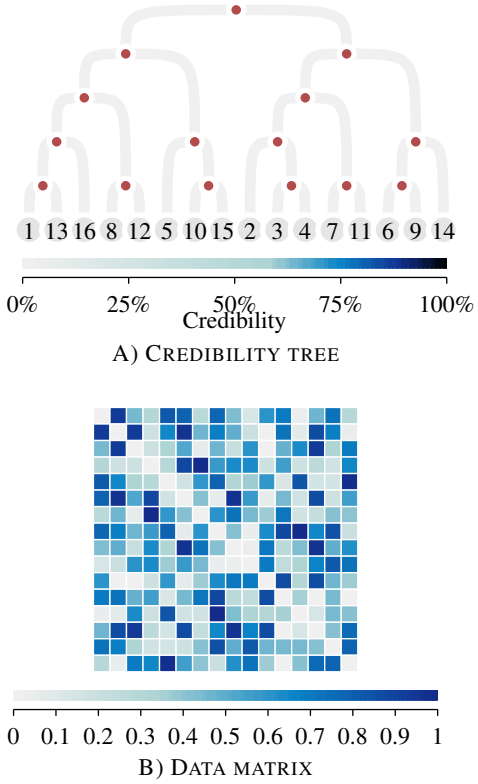


Fig. 3. Analysis of toy example 2 in which there is no structure in the data. The output of the proposed cluster analysis is a hierarchical clustering with very low credibility, indicating that there is no salient hierarchical structure in the data.

2.5. Summarizing the posterior distribution

The output of the analysis is a (posterior) distribution over multifurcating trees. The distribution is represented as a sample of trees as generated by the MCMC procedure. If one is interested in computing quantitative summary statistics, this can be done by computing the statistic for each realization in the posterior sample and averaging.

If we are interested in understanding the hierarchical structure in the data, it can be beneficial to construct a graphical summary of the posterior distribution. One idea could be to plot the most likely tree, i.e. the tree that maximizes the posterior distribution; we denote this the MAP tree. Unfortunately, this only illustrates a single realization and does not provide information about the uncertainty associated with the hierarchical structure, which is quantified by the posterior distribution.

A better graphical summary would illustrate both the most probable hierarchical structure as well as the confidence we have in the structure. To this end, we visualize the tree with the highest posterior credibility and color code the branches according to their credibility. Starting at the root, we find the most likely split according to the posterior distribution and color it according to its probability by computing the proportion of posterior samples which possess this split. For each split below, we again recursively find the most likely split, and color it according to the probability of seeing this particular split and all previous splits. This *credibility tree* makes it easy to visually read off the credibility interval for a particular tree structure.

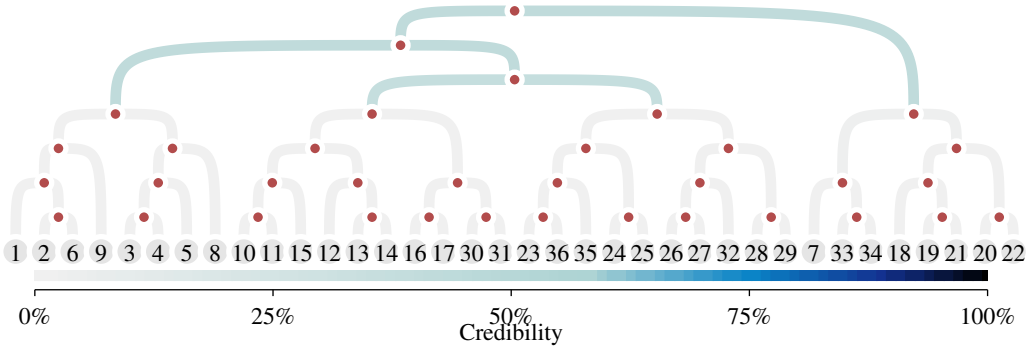
3. EXPERIMENTAL RESULT

In all experiments the hyper-parameters were set to $m = 0$, $k = 1$, $a = 1$, $b = 1$, $\beta = 1$, $\alpha = 0$. For each data set we generated 200,000 samples, discarded the first half for burn-in, and thinned the remaining by a factor of 100 to yield 1,000 posterior samples. Each MCMC analysis was repeated 10 times and the samples were pooled, yielding 10,000 posterior samples as the final result.

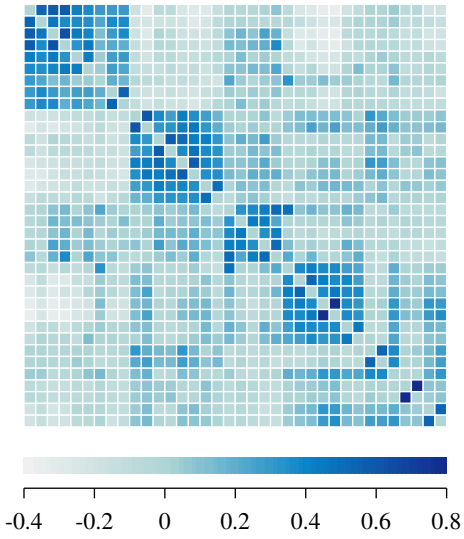
3.1. Demonstration on toy data

For demonstration, we generated two simple toy data examples both with 16 data objects. In the first example the objects were divided into three groups $\{1, \ldots, 4\}$, $\{5, \ldots, 8\}$, and $\{9, \ldots, 16\}$. The relations between these groups were generated from a unit variance Normal distribution with means 0, 5, and -5. The third group was further split into two groups $\{9, \ldots, 12\}$, and $\{13, \ldots, 16\}$ with relations generated with mean 15. Finally, relations within the four groups were generated with means 20, -10, 5, and 5, yielding no further structure in the data. The result of the cluster analysis is given in Fig. 2. The structure in the data is perfectly inferred with almost 100% credibility assigned to the correct structure.

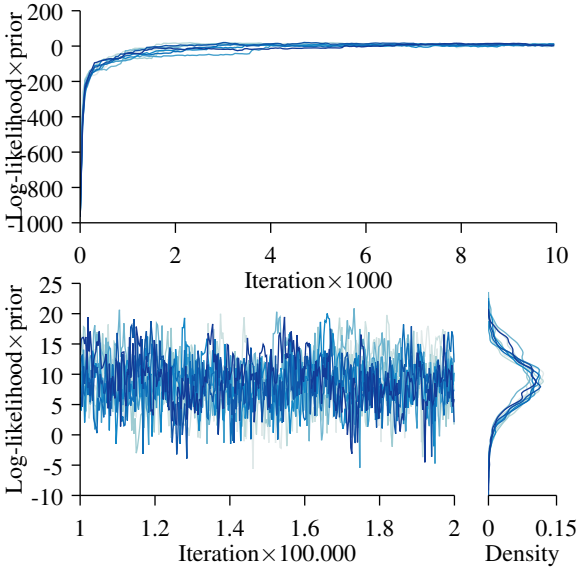
In the second example the relations between objects were generated independently from a uniform distribution. This was done to illustrate what the proposed method would output when there is no explicit structure in the data. The result of the cluster analysis is given in Fig. 3. Here, the credibility tree indicates that there is no salient hierarchical structure in the data, since the most credible solution has very low posterior probability.



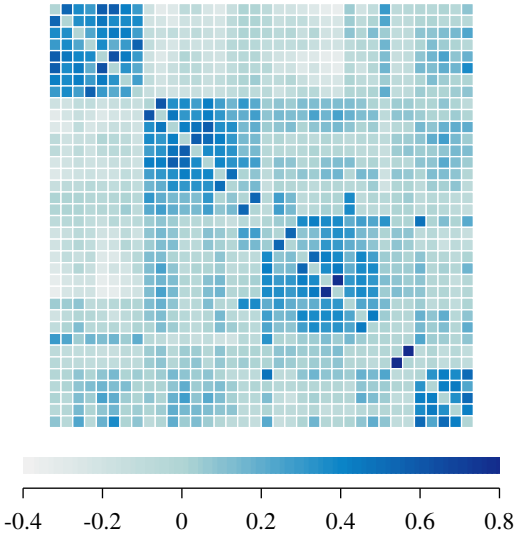
A) CREDIBILITY TREE



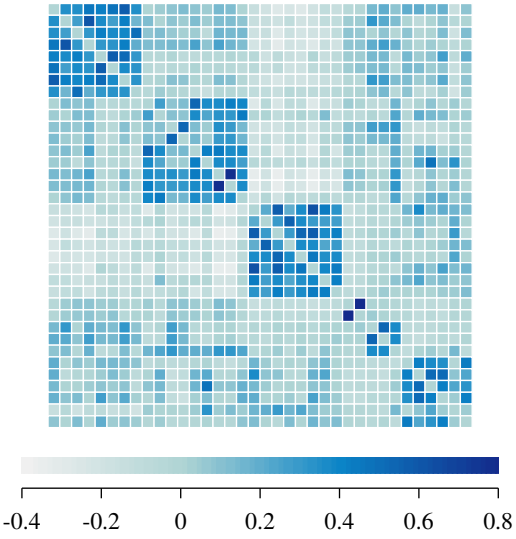
B) BRAIN NETWORK PERMUTATION



C) MCMC CONVERGENCE

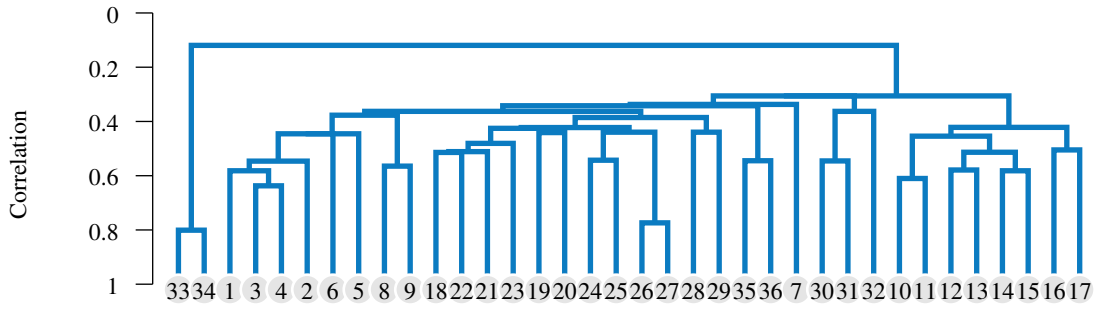


D) CREDIBILITY PERMUTATION

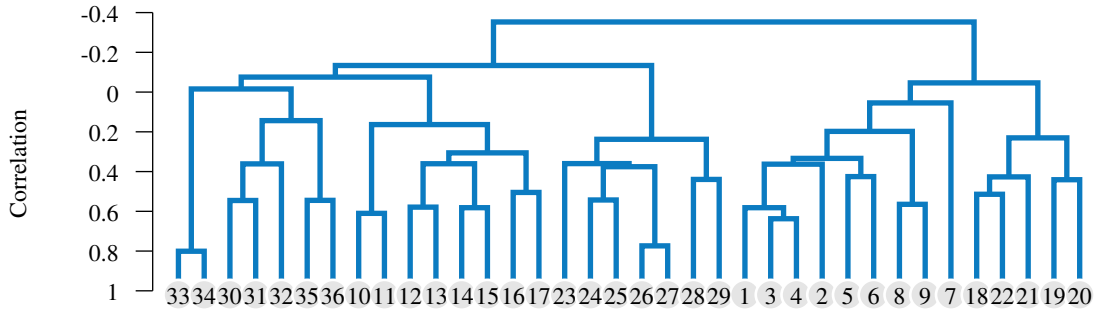


E) MAP PERMUTATION

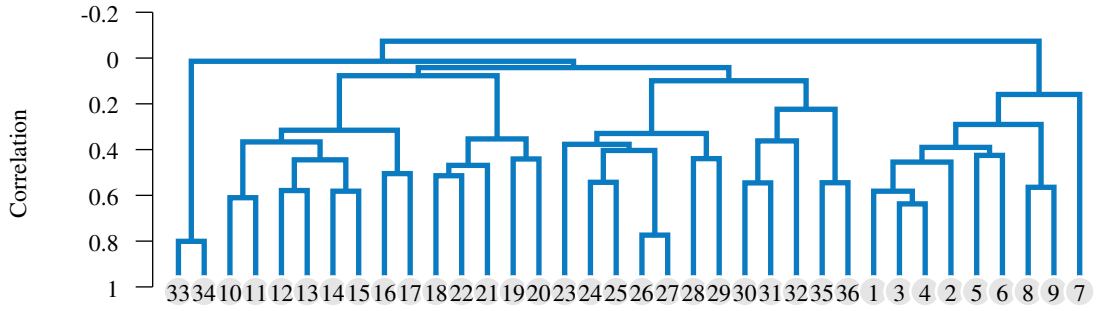
Fig. 4. Analysis of brain network data (I).



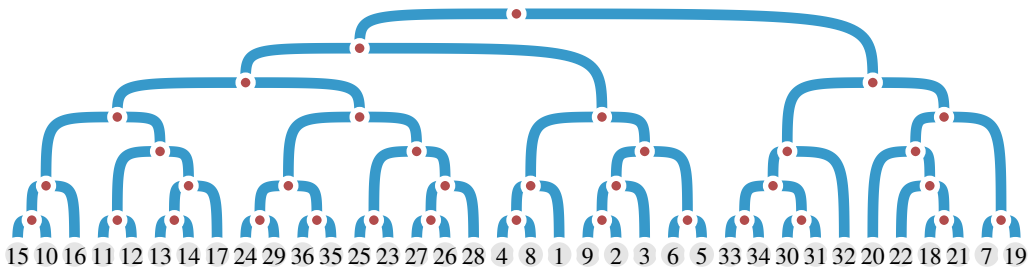
A) HIERARCHICAL CLUSTERING, SINGLE LINKAGE



B) HIERARCHICAL CLUSTERING, COMPLETE LINKAGE



C) HIERARCHICAL CLUSTERING, AVERAGE LINKAGE



B) MAP TREE

Fig. 5. Analysis of brain network data (II).

3.2. Functional Brain Connectivity Data

We applied the proposed method to a data set of correlations between 36 brain regions [12] which reflects 7 brain networks corresponding to data objects $\{1,\ldots,9\}$, $\{10,\ldots,17\}$, $\{18,\ldots,22\}$, $\{23,\ldots,29\}$, $\{30,31,32\}$, $\{33,34\}$, and $\{35,36\}$. When organizing the data matrix according to these brain networks, it clearly reflects a cluster structure (see Fig. 4.B). Marcus E. Raichle states that *"...while correlations within networks appear distinctive in this presentation, relationships among networks (both positive and negative) are also prominent, emphasizing the integrated nature of the brain's functional organization, which is sometimes overlooked ..."* [12]. This integrated nature can be exploited by our current unsupervised modeling approach based on structural similarity.

Fig. 4.C shows the logarithm of the likelihood$\times$prior for the 10 MCMC runs for the first 1,000 and the last 100,000 samples indicating that the sampler converges. A kernel density estimate of the log likelihood$\times$prior computed for the second half of the samples indicates that the MCMC sampler mixes since the distribution is roughly equal for the ten independent chains.

The results in Fig. 4 show that according to our model there are 4 statistically salient clusters roughly equal to the first four brain networks which are also the largest of the seven and therefore those which we would expect to have the most statistical support. The two smallest brain networks are also clustered together in the credibility tree, but with very low credibility. The cluster consisting of objects $\{30,31,32\}$ are not close in the credibility tree, but are close in the MAP tree (see Fig. 5.D).

For comparison we also computed three standard hierarchical clusterings (see Fig. 5.A–C), where the correlation was used as similarity measure. From the single linkage dendrogram in Fig. 5.A it can be seen that the $\{33,34\}$ cluster is very prominent as these two data objects have high correlation with each other but low correlation to everything else. This is exactly the type of structure a direct similarity based clustering is designed to find. All three standard hierarchical clustering methods find clusters corresponding to the larger brain networks, and the complete and average linkage results closely match the known brain networks. It is, however, difficult to assess from the dendrogram alone where to make the best cut-off, and no single cut-off results in a clustering exactly equal to the known brain networks.

A clear difference between the *direct* and *structural* similarity can be seen in the brain network consisting of nodes $\{30,31,32\}$. In the data matrix in Fig. 4.B it is evident that the tree nodes are highly correlated with each other, but in addition to that, 30 and 31 are correlated with the nodes in $\{10,\ldots,17\}$ whereas 32 is correlated with nodes in $\{23,\ldots,29\}$. This is reflected in the solution shown in the credibility tree in Fig. 4.A but not in any of the standard hierarchical clusterings which do not consider structural similarity.

4. DISCUSSION

We have proposed a model for learning hierarchical clustering based on structural similarity. The utility of the current framework when compared to traditional agglomerative hierarchical clustering is that the method is based on a global objective (a generative statistical model) rather than local merge-heuristics, and that the inferred posterior distribution over trees can be used to assess the credibility of the hierarchical structures. In the analysis of the synthetic data our method correctly identified the prominent underlying hierarchy giving it a high credibility whereas data without explicit structure also resulted in very low credibility of the inferred hierarchies. For the functional brain connectivity data the method was able to identify the largest brain network with high credibility by exploiting the structural similarity of the data. Furthermore, the proposed method suffered less from the problem of choosing an appropriate cut-off level as in agglomerative clustering.

For the considered problem the sampler was able to adequately mix. However, we expect mixing to be an issue when analyzing larger systems. Furthermore, a limitation of the proposed method is that sampling hundreds of thousands of trees is inherently slower than constructing a single dendrogram. Thus, future work should focus on constructing efficient inference procedures. In particular, a key advantage of a hierarchical model is its inherent nested structure that admits parallel computing [8].

5. REFERENCES

[1] John A Hartigan and Manchek A Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[2] Carl Edward Rasmussen, "The infinite gaussian mixture model.," in *NIPS*, 1999, vol. 12, pp. 554–560.

[3] Katherine a. Heller and Zoubin Ghahramani, "Bayesian hierarchical clustering," in *Proceedings of the 22nd international conference on Machine learning - ICML '05*. 2005, pp. 297–304, ACM Press.

[4] Stanley Wasserman and Carolyn Anderson, "Stochastic a posteriori blockmodels: Construction and assessment," 1987.

[5] Katherine Faust and Stanley Wasserman, "Blockmodels: Interpretation and evaluation," 1992.

[6] Aaron Clauset, Cristopher Moore, and Mark EJ Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[7] Tue Herlau, Morten Mørup, Mikkel Nørgaard Schmidt, and Lars Kai Hansen, "Detecting hierarchical structure in networks," in *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*. IEEE, 2012, pp. 1–6.

[8] Tue Herlau, Morten Mørup, and Mikkel N Schmidt, "Modeling temporal evolution and multiscale structure in networks," in *JMLR W&CP*, 2013, vol. 28(3), pp. 960–968.

[9] Peter McCullagh, Jim Pitman, and Matthias Winkel, "Gibbs fragmentation trees," *Bernoulli*, vol. 14, no. 4, pp. 988–1002, 2008.

[10] Mikkel N. Schmidt, Tue Herlau, and Morten Mørup, "Nonparametric bayesian models of hierarchical structure in complex networks," arXiv:1311.1033, 2013.

[11] Joseph Felsenstein, "The number of evolutionary trees," *Systematic Zoology*, vol. 27, no. 1, pp. 27–33, Dec. 1978.

[12] Marcus E. Raichle, "The Restless Brain," 2011.