Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Information Systems 29 (2004) 315–326

www.elsevier.com/locate/infosys

Hierarchical model-based clustering of large datasets through fractionation and refractionation

Jeremy Tantrum*[,1], Alejandro Murua, Werner Stuetzle[2]

Department of Statistics, University of Washington, Seattle, WA 98195, USA

Abstract

The goal of clustering is to identify distinct groups in a dataset. Compared to non-parametric clustering methods like complete linkage, hierarchical model-based clustering has the advantage of offering a way to estimate the number of groups present in the data. However, its computational cost is quadratic in the number of items to be clustered, and it is therefore not applicable to large problems. We review an idea called Fractionation, originally conceived by Cutting, Karger, Pedersen and Tukey for non-parametric hierarchical clustering of large datasets, and describe an adaptation of Fractionation to model-based clustering. A further extension, called Refractionation, leads to a procedure that can be successful even in the difficult situation where there are large numbers of small groups.

Keywords: Model-based clustering; Fractionation; Refractionation

1. Introduction

The goal of clustering is to identify distinct groups in a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset R^m$. For example, when presented with (a typically higher dimensional version of) a dataset like the one in Fig. 1 we would like to detect that there appear to be (perhaps) five or six distinct groups, and assign a group label to each observation. (Throughout this paper we distinguish between "groups" and "clusters". The latter are estimates of the groups.)

To cast clustering as a statistical problem we regard the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as a sample from some unknown probability density $p(\boldsymbol{x})$. There are two statistical approaches to clustering. *Non-parametric clustering* [1–4] is based on the premise that groups correspond to modes of the density $p(\boldsymbol{x})$. The goal then is to estimate the modes and assign each observation to the "domain of attraction" of a mode. In contrast, *model-based clustering* [5–8] assumes that each group g is represented by a density $p_g(\boldsymbol{x})$ that is a member of some parametric family, such as the multivariate normal family. The density $p(\boldsymbol{x})$ then is a mixture of the group densities, and the parameters of the mixture components as well as their number can be

*Corresponding author.
E-mail addresses: tantrum@stat.washington.edu (J. Tantrum), murua@ieee.org (A. Murua), wxs@stat.washington.edu (W. Stuetzle).
[1] Supported by NSA contracts 62-1942 and 62-2948.
[2] Supported by NSF grant DMS-9803226 and NSA contracts 62-1942 and 62-2948.

0306-4379/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4379(03)00073-5

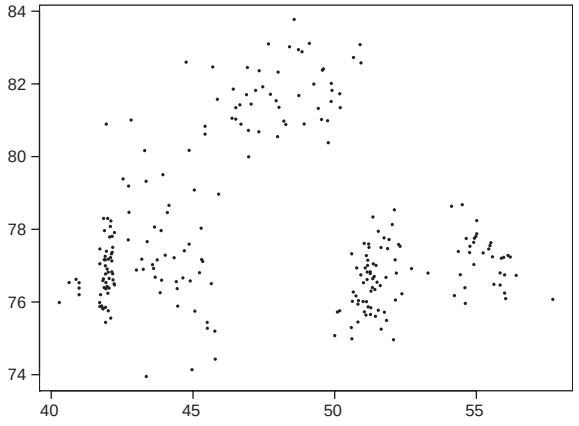

Fig. 1. Dataset with 5–6 apparent groups.

estimated from the data. The ability to estimate the number of groups is an important strength of the model-based approach. There is, as yet, no comparable method for non-parametric clustering in more than one dimension.

In this paper we focus on model-based clustering of large datasets. We review an idea called Fractionation, originally conceived by Cutting et al., [9] for non-parametric hierarchical clustering of large datasets, and describe an adaptation of Fractionation to model-based clustering. A further extension, called Refractionation, leads to a procedure that can be successful even in the difficult situation where there are large numbers of small groups. Refractionation is the principal new idea presented in our paper.

1.1. Model-based clustering in a nutshell

The underlying assumption of model-based clustering is that the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are a sample from a mixture density $p(\boldsymbol{x}) = \sum_{g=1}^{G} \pi_g \, p_g(\boldsymbol{x})$. Here, π_g is the prior probability that a randomly chosen observation belongs to group g, and p_g is the density modeling group g. A popular model, and the one we focus on, is to assume that the group densities p_g are multivariate Gaussian with mean μ_g and covariance matrix Σ_g.

The log-likelihood of the data for a Gaussian mixture with a given number G of mixture components is

$$L = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} \pi_g \, \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \Sigma_g) \right), \tag{1}$$

where $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. This log-likelihood can be optimized over the $\pi_g, \boldsymbol{\mu}_g$, and Σ_g using the EM-algorithm [6, Chapter 2.8.].

There are many ways of estimating the number G of groups, or components, in a mixture model [6, Chapter 6]. For example, we can find the number $\hat{G}$ that maximizes the Bayesian information criterion (BIC) [10,11]:

$$\hat{G} = \arg\max_{G}(2 \times L(G) - r \log(n)). \tag{2}$$

Here, $L(G)$ is the log-likelihood of the best G component model, r is the number of parameters of the model and n is the number of observations.

While attractive conceptually, this approach to fitting mixture models and estimating the number of components is slow, because it requires optimizing the log-likelihood for many different values of G. Following Fraley and Raftery [11], we address this problem by using a hierarchical approach: Find a model with $G-1$ components by merging the two groups of the G component model for which the merge leads to the smallest decrease in log-likelihood.

Use of the BIC for choosing the number of mixture components is only justified if the mixture models are fitted by maximum likelihood. Strictly speaking this requires running the EM algorithm to convergence for each value of G, which would be computationally expensive and would destroy the hierarchical structure of the partitions that we obtain. As a compromise, we perform an E-step, followed by an M-step and use the resulting mixture likelihood in the BIC formula.

Straight forward implementation of hierarchical model-based clustering leads to an algorithm with complexity at least $O(n^2)$. In contrast, the algorithms presented in Sections 2 and 3 are linear in the number of observations.

1.2. Previous work on model-based clustering for large datasets

There are several ways of extending model-based clustering to large datasets. The simplest and potentially fastest is to draw a sample of the data, fit a mixture model to the sample, and then use Bayes' rule to assign the remaining observations to the clusters. A problem with this approach is that small groups will be represented in the sample by very few observations or be missed altogether. Therefore, the corresponding clusters will be either ill determined or absent.

Another method of fitting mixture models to large datasets is the Scalable EM (SEM) algorithm of Bradley et al., [12,13]. Their method requires only a single scan of the dataset. Its main drawback is that it does not offer a way of estimating the number of groups or mixture components; the number of clusters is a parameter of the procedure.

Domingos and Hulten's [14] approach is similar to the one proposed in [12,13]. They cluster the data in manageable section and pass through the dataset only once. The biggest difference is that Domingos and Hulten assume that they work on an infinite data stream and so choose to stop when their estimates of the clusters are not changing significantly. The number of clusters is a parameter of the procedure.

2. Fractionation

Fractionation was originally presented by Cutting et al., [9] as a method for extending $O(n^2)$ hierarchical clustering methods to large datasets. In their application the desired number G of clusters was specified a priori; there was no attempt at estimating the number of groups in the data. Let M be the largest number of items to which we can reasonably apply the base hierarchical clustering procedure.

The original Fractionation algorithm proceeds as follows:

1. Split the data into subsets or fractions of size M.
2. Cluster each fraction into a fixed number αM of clusters, with $\alpha < 1$. Summarize each cluster by its mean. We refer to these cluster means as *meta-observations*.
3. If the total number of meta-observations is greater than M, return to step (1), with the meta-observations taking the place of the original data.
4. Cluster the meta-observations into G clusters.
5. Assign each individual observation to the cluster with the closest mean.

The number of fractions in the ith iteration is $\alpha^{i-1}n/M$ and the work involved in clustering a fraction is $O(M^2)$ independent of n. This shows that the total run time is linear in n and decreasing in α.

2.1. Model-based Fractionation

If we use hierarchical model-based clustering as the base clustering method in Fractionation, then we get model-based Fractionation. The main difference between the Fractionation method of Cutting et al. [9] and model-based Fractionation is that in model-based Fractionation a meta-observation is characterized not just by a mean, but by all the sufficient statistics, i.e. the mean, the covariance, and the number of observations in the cluster.

We do not want to assume that the number of groups is known a priori. Instead, we determine the number of clusters (mixture components) in Step 4 of the Fractionation algorithm using the BIC.

3. Model-based refractionation

A major problem with Fractionation is that once observations from different groups have been assigned to the same meta-observation this error will never be corrected. Such erroneous assignments are less likely to occur if fractions are pure, i.e. contain observations from few groups or, equivalently, if groups are split over few fractions. We could form purer fractions if we knew the group labels of the observations. This observation

suggests applying Fractionation repeatedly and forming the fractions for Step 1 of the ith pass based on the clustering produced in the $(i-1)$st pass. Conceptually, Step 4 of the Fractionation algorithm is replaced by two steps, both involving hierarchical model-based clustering of the meta-observations generated by Step 3:

4a. Cluster the meta-observations into G clusters, where G is determined by the BIC.
4b. Define the fractions for the ith pass: as soon as a cluster formed during the merging represents more than M observations, make those observations into a fraction and remove the cluster from the merge process.

We stop the Refractionation iterations when the similarity between consecutive partitions no longer increases.

3.1. Illustration

To illustrate how Refractionation works, consider a simple example in two dimensions with 25 equally spaced Gaussian groups containing 16 points each. Fig. 2 shows the data and the component densities of the model. The circles in this and the following figures are isopleths of the component densities containing 95% of the mass.

We randomly split the data into four fractions of 100 observations each (Step 1 of the Fractionation algorithm), and then use model-based hierarchical clustering to cluster each fraction into $M/10 = 10$ clusters (Step 2 of the algorithm). The fractions and their clusters are shown in Fig. 3.

The number of meta-observations produced by clustering the fractions in this case is 40 which is less than $M = 100$ (Step 3) and we can therefore proceed to Steps 4a and 4b.

Clustering the 40 meta-observations into 25 clusters (Step 4a) produces the mixture model whose component densities are shown in Fig. 4. Clearly, this clustering in no way reflects the structure of the data.

Clustering the 40 meta-observations into new fractions (Step 4b) results in fraction sizes of 97, 108, 104, and 91. Fig. 5 shows the new fractions.

We now start the second pass of Fractionation. Each fraction again is clustered into 10 clusters (Step 2) shown in Fig. 5.

Clustering the 40 meta-observations into 25 clusters (Step 4a) produces the mixture model shown in Fig. 6. We have essentially recovered the structure of the data.

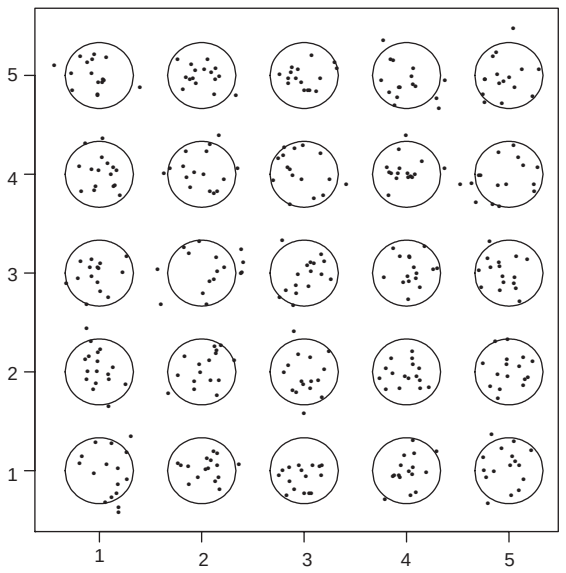

Fig. 2. Observations and component densities.

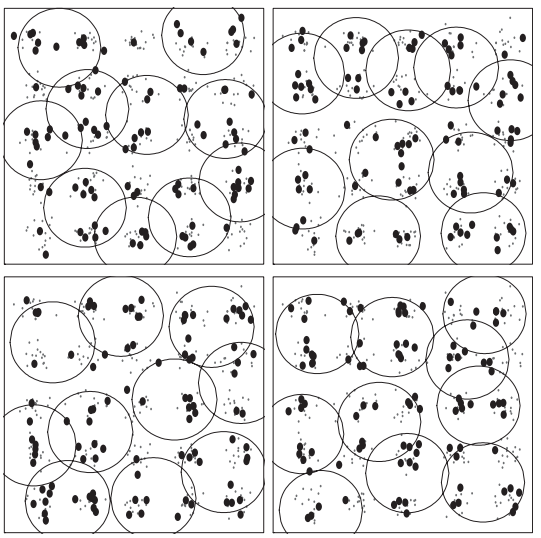

Fig. 3. Meta-observations obtained by clustering the initial four fractions.

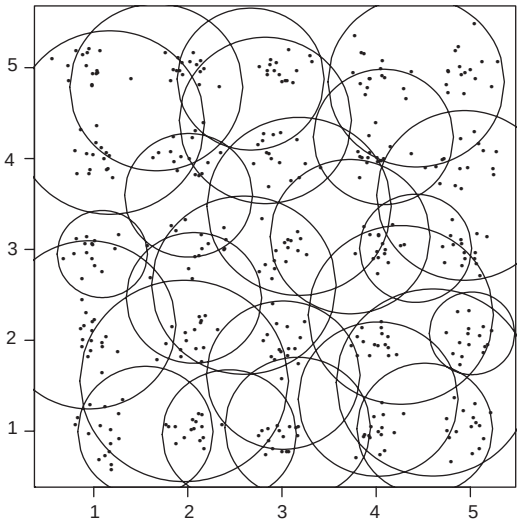

Fig. 4. Clusters after the first pass of Fractionation.

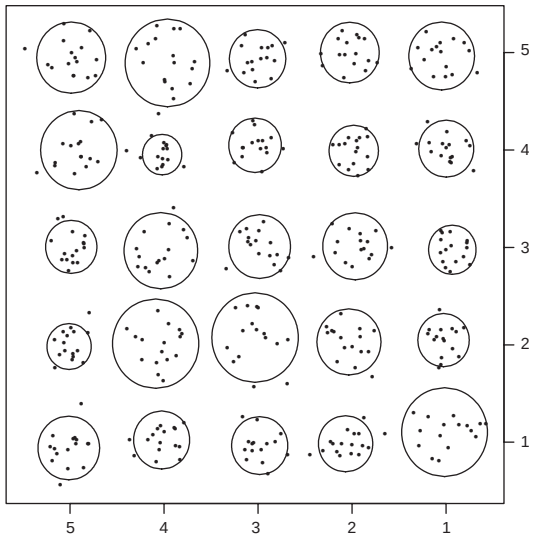

Fig. 6. Clusters after the second pass of Fractionation.

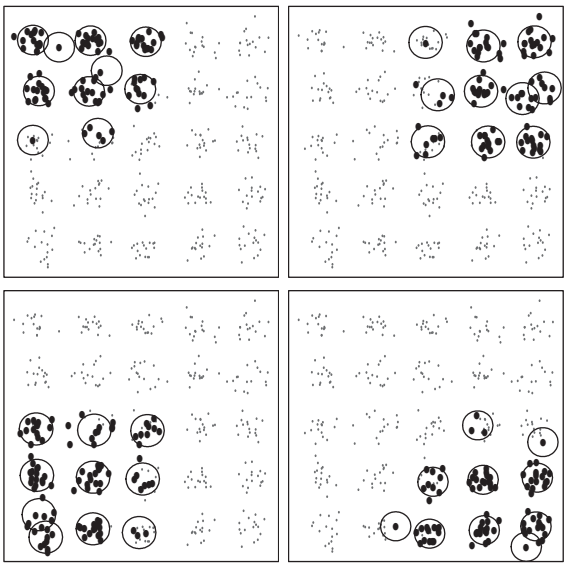

Fig. 5. Meta-observations obtained by clustering the four fractions in the second pass of Fractionation.

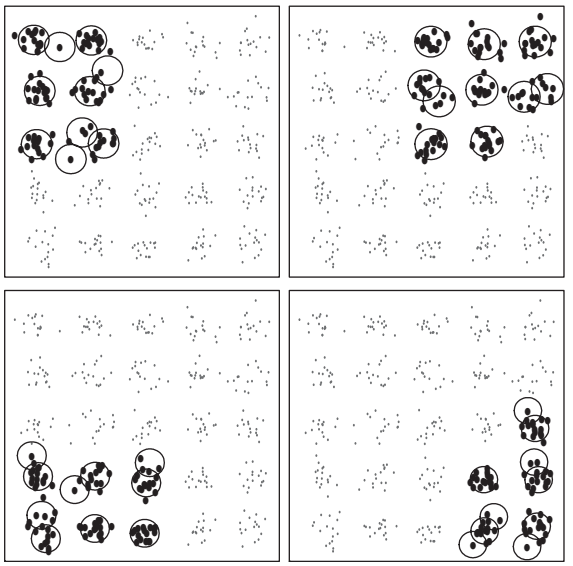

Fig. 7. Meta-observations obtained by clustering the four fractions in the third pass of Fractionation.

A third pass of Fractionation (Figs. 7 and 8) leads to almost the same mixture model (Fig. 8) as the second pass (Fig. 6), and the Refractionation process stops.

Table 1 gives numerical summaries of the purity of the fractions. At the beginning of the first Fractionation pass, each of the 25 groups is scattered over all four fractions, whereas at the beginning of the third pass only one of the groups is split across multiple fractions.

3.2. Scope of (Re)Fractionation

In order to gain some insight into the scope and limitations of (Re)Fractionation, we consider an idealized situation where the groups are so well separated that it is unambiguous whether or not two observations or meta-observations belong to the same group. This allows us to separate performance of the base clustering method from the performance of Fractionation and Refractionation.

Let n_g be the number of groups in the data, let n_f be the number of fractions, and let n_c be the number of clusters generated from each fraction in Step 2 of the Fractionation algorithm. Clearly, if $n_g \leqslant n_c$ then Fractionation will work and Refractionation is unnecessary. On the other hand, if $n_g > n_c$ then it is possible for a fraction to contain observations from more than n_c groups, which will lead to impure clusters, and therefore the groups will not be recovered perfectly.

Even in our simple scenario it is difficult to make such simple statements about Refractionation. We can only prove that Refractionation works for $n_g = n_c + 1$, under some restrictions about the group sizes. However, our examples show that the range of applicability is much larger. It is also clear that Refractionation will not recover the groups if $n_g > n_f n_c$. In this case there must be at least one fraction that contains observations from more than n_c groups, and clustering this fraction will lead to impure clusters.

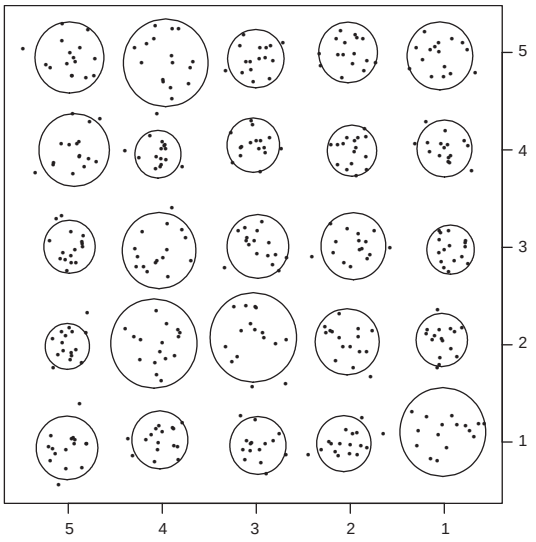

Fig. 8. Clusters after the third pass of Fractionation.

Table 1
The distribution of the number of fractions over which observations from each group are scattered, at the start of each Fractionation pass

Pass	Min	Median	Max	>1	>2
1	4	4	4	25	25
2	1	1	2	10	0
3	1	1	2	1	0

Columns labeled Min, Median, and Max show the minimum, median, and maximum of this distribution, respectively. The last two columns show the number of groups scattered over more than one fraction (>1), and over more than two fractions (>2) at the start of each Fractionation pass.

4. Examples

In order to investigate how well model-based Fractionation and Refractionation can find groups in a dataset, we apply them to four datasets for which the group labels are known.

4.1. Measuring the agreement between groups and clusters

In our examples we know the true group labels of the observations, and we want to measure the degree of agreement between the groups and their estimates, the clusters. We use the Fowlkes–Mallows index [15] as a measure of agreement. The index is the geometric mean of two probabilities: the probability that two randomly chosen observations are in the same cluster given that they are in the same group, and the probability that two randomly chosen observations are in the same group given that they are in the same cluster. Hence a Fowlkes–Mallows index near 1 means that the clusters are a good estimate of the groups.

To compute the Fowlkes–Mallows index we construct a contingency table of the groups and the clusters, as shown in Table 2. Ideally, this table should contain a few non-zero entries (i.e. a few non-empty cells), so that groups are only spread

Table 2
Comparison between clusters (columns) and groups (rows)

True groups	clusters 1	2	...	J	Total
1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
...	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	n

Each cell count n_{ij} is the number of common elements in cluster j and group i.

on a small handful of representative clusters. Let $n_{i\cdot}$ be the sum over the ith row of the table, and let $n_{\cdot j}$ be the sum over the jth column. Then the Fowlkes–Mallows index is given by:

$$\sum_{i,j} \binom{n_{ij}}{2} \Big/ \sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}. \tag{3}$$

4.2. The TDT dataset

We used the Topic detection and tracking document collection [16] for our examples. The documents are news stories from Reuters and CNN for the year from July 1st 1994–June 30th 1995, and they are ordered according to the time and date of the story. There are 15,863 documents in 25,431 dimensions. 25 topics were chosen and all documents associated to these topics were manually labeled—there are 1,131 labeled documents in the collection. Of the 25 topics, 6 have less than 8 documents each. Examples 1–4 use the 1100 documents which belong to the remaining 19 topics. Example 5 uses the entire TDT dataset.

We relied on standard document retrieval technology to convert the documents into vectors. We assembled the term-document matrix, applied the log-Idf transformation to the term counts as suggested by Dumais [17], and then reduced the dimensionality by latent semantic indexing [18]. We reduced the 1100 documents into a 50-dimensional space and the 15863 documents into a 100-dimensional space.

Model-based clustering of the 1100 documents (more precisely, the 1100 50-dimensional vectors corresponding to the documents) into 19 clusters resulted in a Fowlkes–Mallows index of 0.65 and a 19×19 contingency table (analogous to Table 2) with 34 non-zero entries. This is the standard against which we measure the results in the following examples.

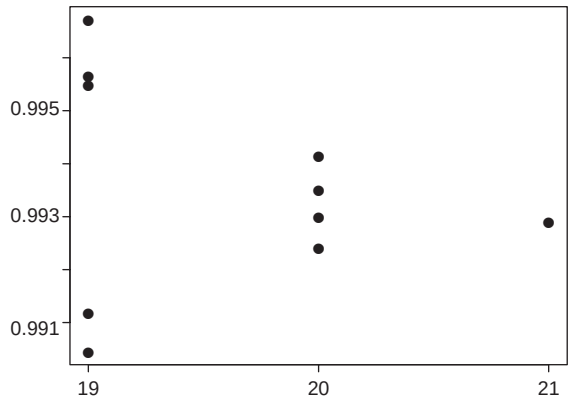

Fig. 9. Fowlkes–Mallows index (vertical axis) vs. number of clusters chosen by the BIC for Example 1.

4.3. Example 1

To create the data for this example, we estimated the mean vector and covariance matrix for each of the 19 groups in the TDT dataset. We then generated 20 times the number of observations in each group from a Gaussian distribution with the group mean vector and covariance matrix. This gave a dataset with $n = 22,000$ observations. We randomly partitioned the data into 22 fractions of $M = 1000$ observations each, and clustered the fractions into $M/10 = 100$ clusters. As the number of groups (19) is small relative to the number of clusters generated in each fraction, one pass of Fractionation was sufficient; no Refractionation was needed. We determined the final number of clusters using the BIC. We repeated this experiment—generating a sample of size 22,000 and clustering it—10 times. In these 10 replications, the BIC chose 19 clusters 5 times, 20 clusters 4 times and 21 clusters once. The average Fowlkes–Mallows index for the runs that chose 19 clusters was 0.9955 and for those that chose 20 it was 0.9932 (see Fig. 9), indicating almost perfect

agreement between groups and clusters. This is reassuring—after all, the data were generated from a Gaussian mixture, and we would hope that model-based clustering would do well.

4.4. Example 2

The data in this example were obtained by estimating each group density by a kernel density estimate [19] and then sampling from this estimate, again generating 20 times the number of observations in the group. We used a Gaussian kernel with the same covariance matrix as the corresponding group scaled by a factor of $1/10$. As in Example 1 this resulted in a dataset of 22,000 observations, but the data are no longer sampled from a mixture of 19 Gaussians. We clustered the dataset using one pass of Fractionation and the BIC for choosing the final number of clusters. In 10 replications of the experiment the BIC chose 21 clusters once, 22 clusters 3 times and 23 clusters 6 times. The values of the Fowlkes–Mallows index were between 0.83 and 0.94 (see Fig. 10).

4.5. Example 3

Examples 1 and 2 are easy: the number of groups is small, and all the groups are large. They could certainly have been recovered by clustering a random sample of manageable size. Example 3 is more challenging.

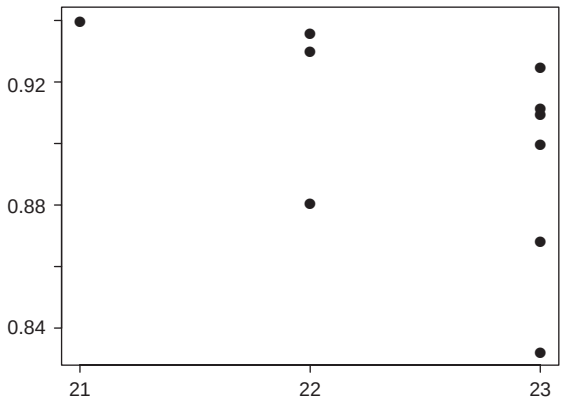

Fig. 10. Fowlkes-Mallows index (vertical axis) vs number of clusters chosen by the BIC for Example 2.

We generated the data for Example 3 by essentially replicating the labeled TDT dataset 19 times, replacing each group by a scaled and shifted version of the entire dataset: Let $\boldsymbol{\mu}_i$ and Σ_i be the mean vector and covariance matrix of the ith group. We obtained the ith replicate by scaling and shifting the entire dataset to have mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i. We end up with $19 \times 19 = 361$ groups and $19 \times 1100 = 20,900$ observations.

We randomly split these 20,900 observations into $M = 20$ fractions of 1045 observations each and clustered fractions into 100 clusters. Because the number of groups (361) is larger than the number of clusters per fraction (100), and initial fractions typically contain observations from more than 100 groups, a single pass through Fractionation does not result in a good clustering of the data, and Refractionation is necessary.

Table 3 shows the Fowlkes–Mallows index of the clustering into 361 clusters after the first four passes through Fractionation. The index almost doubles, indicating that the agreement between groups and clusters improves dramatically. This improvement goes along with an equally drastic decrease in the number of non-zero entries in the 361×361 contingency table.

Tables 4 and 5 confirm that Refractionation indeed increases the purity of the fractions. Table 4 shows that, initially, groups are scattered over many fractions, while after the fourth pass through Fractionation 320 of the 361 groups are contained entirely in a single fraction, and the remaining 41 groups are each split across two fractions.

Table 3
Example 3: Agreement between clusters and groups after each Fractionation pass

Pass	Fowlkes Mallows	non–zero entries
1	0.325	1729
2	0.554	908
3	0.616	671
4	0.613	651

The second column is the number of non-empty cells in the corresponding contingency table. See text for more details.

Table 4
Example 3: Distribution of the number of fractions over which observations from each group are scattered, at the start of each Fractionation pass

Pass	Min	Median	Max	>1	>2
1	6	18	20	361	361
2	1	4	10	350	287
3	1	1	3	68	7
4	1	1	2	41	0

Columns labeled Min, Median, and Max show the minimum, median, and maximum of this distribution, respectively. The last two columns show the number of groups scattered over more than one fraction (>1), and over more than two fractions (>2) at the start of each Fractionation pass.

Table 5
Example 3: Distribution of the number of groups represented in each fraction at the start of each Fractionation pass

Pass	Min	Median	Max	n_f	$361/n_f$
1	270	289	296	20	18.0
2	18	88	150	18	20.1
3	18	19	60	17	21.2
4	19	19	58	16	22.6

Columns labeled Min, Median, and Max, show the minimum, median, and maximum of this distribution, respectively. The last two columns show the number of fractions n_f, and the number of groups per fraction, respectively, at the start of each Fractionation pass.

Table 5 gives the number of groups represented in each fraction at the beginning of each Fractionation pass. At the beginning of the first pass the least diverse fraction contains observations from 270 groups, and the most diverse fraction contains observations from 296 groups. The median number of groups per fraction is 289. In contrast, at the beginning of the fourth Fractionation pass the least diverse fraction contains observations from 19 groups, and the most diverse fraction contains observations from 58 groups. The median number of groups per fraction is 19. These numbers again demonstrate how successful Refractionation is at purifying the fractions. There is no change in clustering after the 4th run of Fractionation.

The groups in this example are too small to fit a mixture model whose components have unconstrained covariance matrices: in 50 dimensions we need at least 51 observations to obtain a non-singular sample covariance matrix. In order to avoid this problem, we constrained the covariance matrices of the mixture components to be diagonal. The results above indicate that this works well if we make the algorithm produce the correct number of clusters, 361. The BIC, however, chooses too many clusters (approximately 480), since the actual shapes of the groups are not well approximated by axis-parallel ellipsoids.

Fitting accurate and parsimonious mixture models to datasets with many small groups requires Mixtures of Factor Analyzers [6,20], combined with a criterion like the BIC to estimate the number of components. As we have yet to implement this approach, we shall instead consider a simulated example where the group covariance matrices are diagonal, and therefore a mixture model with diagonal covariance matrices will fit well.

4.6. Example 4

We generated the data for this example in a way very similar to Example 3. However, we replaced the observations in each of the 361 groups by simulated data from a multivariate Gaussian distribution with the same mean as the group and a diagonal covariance matrix obtained by setting the off-diagonal elements of the sample covariance matrix to zero.

We simulated 10 datasets of size 20,900 from this mixture distribution with 361 axis-parallel components and clustered them with the same algorithm as in Example 3, except that we chose the number of clusters using the BIC. In ten replications of this experiment, the number of clusters chosen by the BIC was between 371 and 379, with values of the Fowlkes–Mallows index between 0.781 and 0.806 (see Fig. 11). The Fowlkes–Mallows index for models with 361 clusters is between 0.797 and 0.820, which is only slightly better than for the models obtained using the BIC.

4.7. Example 5

For this example we clustered the entire set of 15,863 TDT documents in 100-dimensional space.

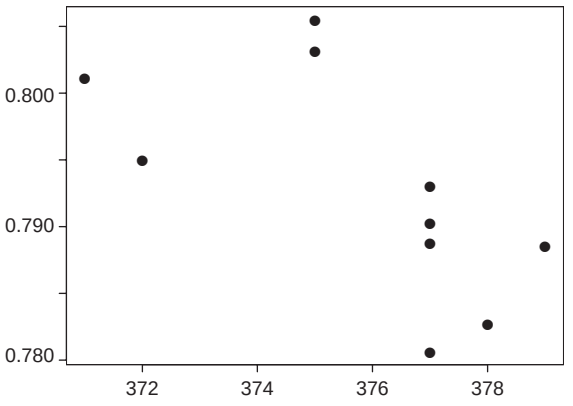

Fig. 11. Fowlkes-Mallows index (vertical axis) vs number of clusters chosen by the BIC for Example 4.

The documents are ordered according to the time and date of the story. We used this ordering to split the data into 15 fractions of approximately 1000 documents each and then applied the Refractionation process as described in Section 3.

Only 1100 of the 15,863 documents are labeled, which raises the issue of measuring the quality of clustering for partially labeled data. Our first suggestion is to use the labeled documents only: each labeled document has both a topic label and a cluster label, and we can measure agreement between these labels using the Fowlkes–Mallows index described in Section 4.1. We call this quality measure the FM1 index.

The results are shown in Table 6. The second column gives the agreement between partitions obtained in successive passes of Fractionation, as measured by the Fowlkes–Mallows index. After 10 passes the agreement no longer improved, which led us to stop the Refractionation process.

The third column shows the FM1 index for successive passes of Fractionation. There is no improvement in the index, which suggests that Refractionation was either unsuccessful or unnecessary. We suspect the latter. Our explanation for the lack of improvement is that documents on the same topic are concentrated in time, and therefore, the initial fractions (obtained by splitting the collection according to time) were already quite pure.

Table 6
Example 5: Agreement between consecutive partitions and FM1 index for the 15,863 TDT documents when using time order to form the original fractions

Pass	Agreement between consecutive partitions	FM1 index
1		0.545
2	0.269	0.516
3	0.475	0.539
4	0.557	0.530
5	0.624	0.492
6	0.626	0.530
7	0.667	0.530
8	0.692	0.535
9	0.711	0.541
10	0.733	0.546
11	0.727	0.545

Table 7
Example 5: Agreement between consecutive partitions and FM1 index for the 15,863 TDT documents when using random assignment to form the original fractions

Pass	Agreement between consecutive partitions	FM1 index
1		0.479
2	0.248	0.500
3	0.528	0.517
4	0.539	0.517
5	0.594	0.526
6	0.625	0.524
7	0.621	0.530
8	0.665	0.555
9	0.665	0.518
10	0.711	0.544
11	0.713	0.569
12	0.626	0.536

To verify this explanation we repeated the experiment, but randomly assigned documents to the initial fractions. The results are shown in Table 7. This time there is a significant increase in the FM1 index over successive passes of Fractionation. The final values of the FM1 index for the two different choices of initial fractions are close, suggesting that the penalty for a poor choice of the initial fractions is small.

A criticism of the FM1 index is that it only reflects "recall" and ignores "precision". Consider the ideal situation where the labeled documents for each topic are concentrated in exactly one cluster

in such a way that labeled documents from different topics fall in different clusters. This situation would result in an FM1 index of 1, no matter how many additional unlabeled documents are contained in these clusters. This suggests an alternative measure of goodness of clustering, which we call the FM2 index. We collect all documents in all the clusters containing labeled observations. If a document in this collection was not assigned to a topic then we label it as "*other*". Each of the documents now has two labels, the topic label (possibly *other*) and the cluster label, and we can measure agreement between them using the Fowlkes–Mallows index. Since the FM2 index uses the same data as the FM1 index plus the additional "topic" *other,* the value of FM2 cannot exceed that of FM1.

The labeled subset of the TDT collection was generated by selecting a list of topics and then identifying all the documents referring to those topics. Consequently, the clusters containing labeled documents should have no unlabeled documents, and the FM2 index should be close to FM1 in value. This turns out not to be the case. The optimal number of clusters according to the BIC is 209, and the FM2 index for the corresponding clustering is only 0.12. Even in clusters containing documents from one of the selected topics those documents form a minority.

We do not yet have a conclusive explanation for this contradiction. Graphical exploration of the data using the XGobi visualization system [21] seems to suggest that the labeled subset is not "typical" for the entire collection. While inspection of the labeled subset quickly reveals obvious clusters corresponding to the topics, this does not hold for the full collection. Of course this might be due to shortcomings in the visualization process, and more work is needed.

5. Conclusions

We have proposed model-based Fractionation and Refractionation, methods for extending the range of model-based hierarchical clustering to datasets with tens of thousands of observations and hundreds of groups. Compared with competing approaches to model-based clustering of large datasets, model-based Refractionation does not require that the number of groups in the data be known a priori; it can be estimated from the data. Initial experiments presented in the paper are encouraging. They provide evidence that the heuristics underlying our method indeed appear to be valid.

There are a number of areas for future work. Most importantly, we want to study the performance of the BIC for estimating the number of groups in a Mixture of Factor Analyzers model, in situations where both the size of the dataset and the number of groups are large. We also plan to extend the scope of model-based Refractionation to problems that are another order of magnitude larger than those tackled here, problems with hundreds of thousands of observations and thousands of groups.

References

[1] D. Wishart, Mode analysis: a generalization of nearest neighbor which reduces chaining effects, in: A. Cole (Ed.), Numerical Taxonomy, Academic Press, New York, 1969, pp. 282–311.

[2] J. Hartigan, Statistical theory in clustering, J. Classification 2 (1985) 63–76.

[3] M. Ankerst, M. Breuning, H. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, in: Proceedings, ACM SIGMOD International Conference on Management of Data (SIGMOD'99), Philadelphia, PA, 1999, pp. 49–60.

[4] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery, Data Mining (KDD-96), Portland, OR, 1996, pp. 226–231.

[5] G. McLachlan, K. Basford, Mixture Models: Inference, Applications to Clustering, Marcel Dekker, New York, 1988.

[6] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[7] J.D. Banfield, A. Raftery, Model-Based Gaussian, non-Gaussian clustering, Biometrics 49 (1993) 803–821.

[8] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, density estimation, J. Amer. Statist. Assoc. 97 (2002) 611–631.

[9] D. Cutting, D. Karger, J. Pedersen, J. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections, in: Proceedings 15th Annual International ACM SIGIR Conference on Research,

Development in Information Retrieval, Copenhagen, Denmark, 1992, pp. 318–329.
[10] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 497–511.
[11] C. Fraley, A. Raftery, How many clusters? which clustering method? answers via model-based cluster analysis, The Comput. J. 41 (8) (1998) 578–588.
[12] P. Bradley, U. Fayyad, C. Reina, Scaling clustering algorithms to large datasets, In: Proceedings of the fourth International Conference on Knowledge Discovery, Data Mining (KDD98), New York, NY, 1998.
[13] P. Bradley, U. Fayyad, C. Reina, Scaling EM (expectation-maximization) clustering to large databases, Technical Report MSR-TR-98-35, Microsoft Research, 1999.
[14] P. Domingos, G. Hulten, Learning from innite data in nite time, in: Advances in Neural Information Processing Systems, Vol. 14, MIT Press, Cambridge, MA, 2002.
[15] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. the Amer. Statist. Assoc. 78 (1983) 553–569.
[16] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic Detection, Tracking Pilot Study, Final Report, 1998.
[17] S. Dumais, Improving the retrieval of information from external sources, Behavior Res. Methods, Instrum. Comput. 23 (2) (1991) 229–236.
[18] M. Berry, S. Dumais, G. O'Brien, Using linear algebra for intelligent information retrieval, SIAM Rev. 37 (4) (1995) 573–595.
[19] D. Scott, Multivariate Density Estimation, Wiley, New York, 1992.
[20] G.E. Hinton, P. Dayan, M. Revow, Modeling the Manifolds of Images of Handwritten Digits, IEEE Trans. Neural Networks 8 (1) (1997) 65–74.
[21] D.F. Swayne, D. Cook, A. Buja, XGobi: interactive dynamic data visualization in the X window system, J. Comput. Graphical Statist. 7 (1998) 113–130.