

Hierarchical Model-Based Clustering of Large Datasets Through Fractionation and Refractionation.

Jeremy Tantrum* Alejandro Murua Werner Stuetzle†
University of *Insightful* *University of*
Washington *Corporation* *Washington*

June 3, 2002

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.

*Supported by NSA grant 62-1942.

†Supported by NSF grant DMS-9803226 and NSA grant 62-194 2.

1 Introduction

The goal of clustering is to identify distinct groups in a dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^m$. For example, when presented with (a typically higher dimensional version of) a dataset like the one in Figure 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”, which are estimates for the groups.)

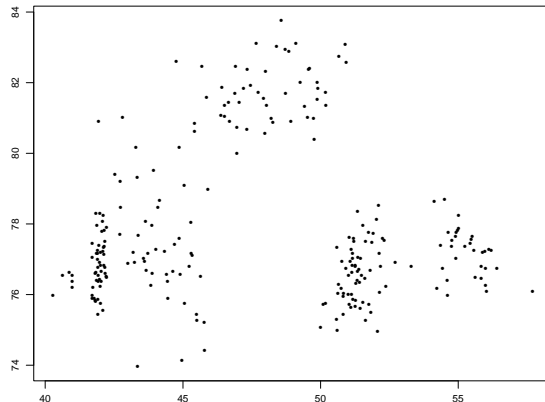


Figure 1: Dataset with 5–6 apparent groups.

To cast clustering as a statistical problem we regard the data $\mathbf{x}_1, \dots, \mathbf{x}_n$ as a sample from some unknown probability density $p(\mathbf{x})$. There are two statistical approaches to clustering. *Nonparametric clustering* [18, 13, 2, 10] is based on the premise that groups correspond to modes of the density $p(\mathbf{x})$. 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* (see [15] and references therein) assumes that each group g is represented by a density $p_g(\mathbf{x})$ that is a member of some parametric family, such as the multivariate normal family. The density $p(\mathbf{x})$ then is a mixture of the group densities, and the parameters of the mixture components as well as their number can be 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 nonparametric 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, Karger, Pedersen and Tukey [7] for nonparametric 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 $\mathbf{x}_1, \dots, \mathbf{x}_n$ are a sample from a mixture density $p(\mathbf{x}) = \sum_{g=1}^G \pi_g p_g(\mathbf{x})$. 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

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

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 ([15], chapter 2.8.)

There are many ways of estimating the number G of groups, or components, in a mixture model ([15], chapter 6.) For example, we can find the number \hat{G} that maximizes the Bayesian Information Criterion (BIC) [16, 12]:

$$\hat{G} = \operatorname{argmax}_G (2 \times L(G) - r \log(n)) . \quad (2)$$

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 a suggestion by Fraley and Raftery [12], 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. As we are not maximizing the mixture likelihood (1), use of the BIC for estimating the number of components is not justified; instead we estimate the number of groups by maximizing the Approximate Weight of Evidence [3]:

$$\hat{G} = \operatorname{argmax}_G (2 \times L(G) - 2r (3/2 + \log(n))) . \quad (3)$$

The hierarchical approach can be expected to work well if the groups are clearly separated.

Unfortunately, straightforward implementation of hierarchical model-based clustering leads to an $O(n^2)$ algorithm. 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, Fayyad and Reina [5, 6]. Their method requires only a single scan of the data set. 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 [8] approach is similar to the one proposed in [5, 6]. They cluster the data in manageable sections 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, Karger, Pedersen, and Tukey [7] 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 i -th 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 be-

tween the Fractionation method of Cutting et al.[7] and model-based Fractionation is that in model-based Fractionation a meta-observation is not characterized 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 AWE criterion.

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 i -th 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 AWE criterion.
- 4b** Define the fractions for the i -th 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 change in the number G of clusters and the cluster compositions is small enough.

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. Figure 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 Figure 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.

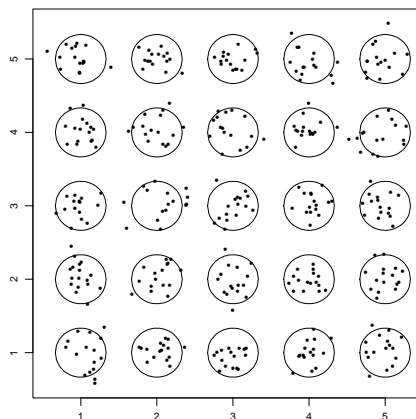


Figure 2: Observations and component densities.

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

Table 1: The distribution of the number of fractions each group resides in at the start of each Fractionation pass.

Clustering the 40 meta-observations into 25 clusters (Step 4a) produces the mixture model whose component densities are shown in Figure 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. Figure 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 Figure 5.

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

A third pass of Fractionation (Figures 7 and 8) leads to almost the same mixture model (Figure 8) as the second pass (Figure 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.

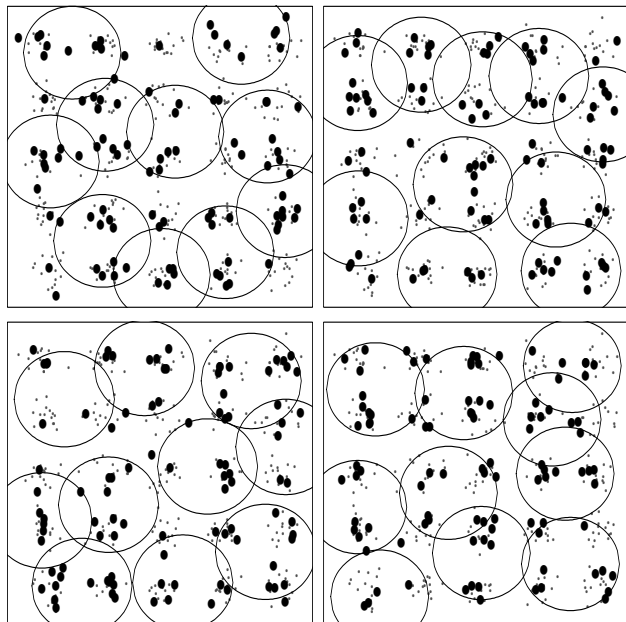


Figure 3: Meta-observations obtained by clustering the initial four fractions.

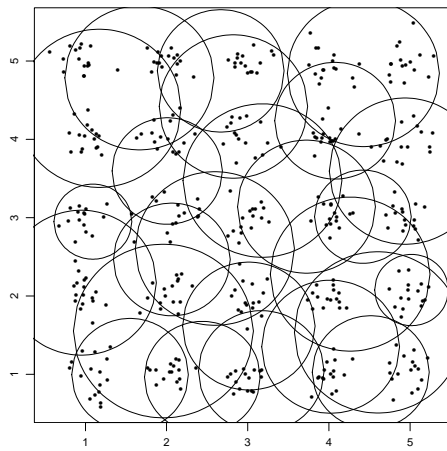


Figure 4: Clusters after the first pass of Fractionation.

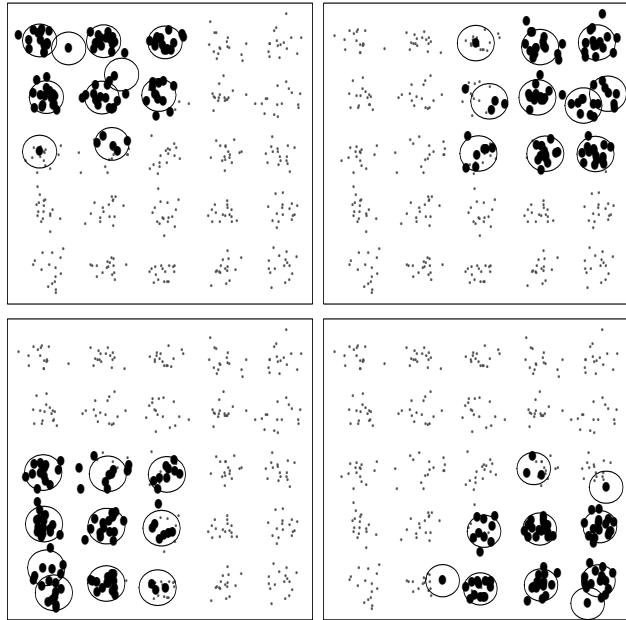


Figure 5: Meta-observations obtained by clustering the four fractions in the second pass of Fractionation.

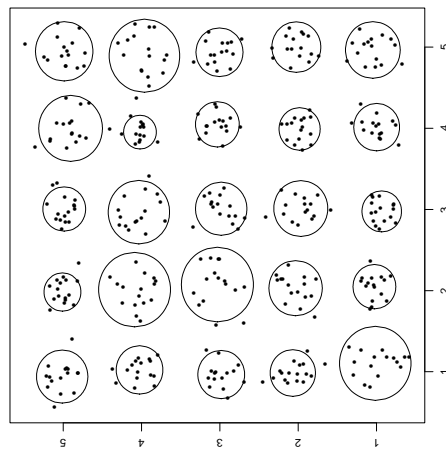


Figure 6: Clusters after the second pass of Fractionation.

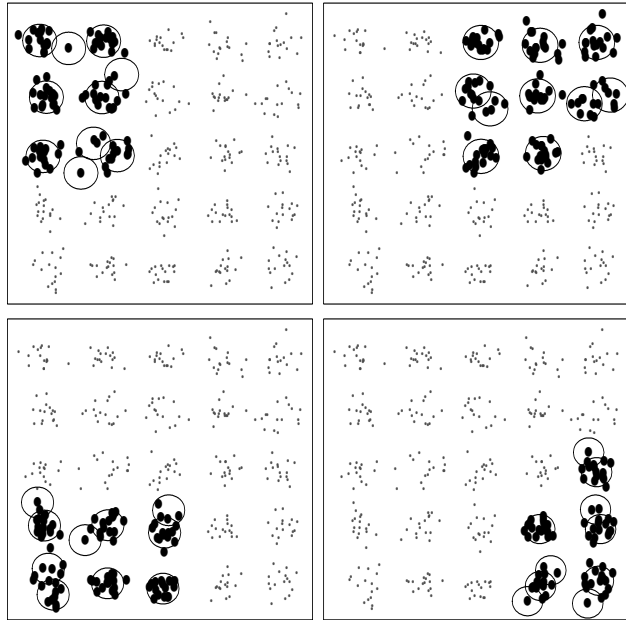


Figure 7: Meta-observations obtained by clustering the four fractions in the third pass of Fractionation.

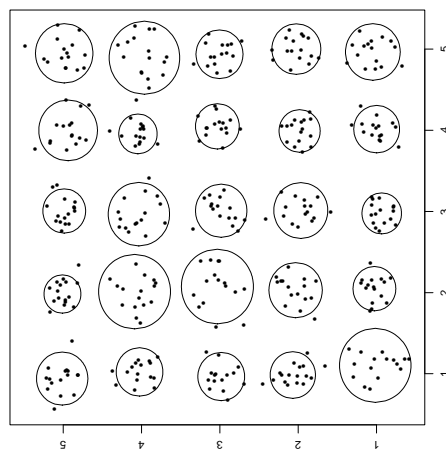


Figure 8: Clusters after the third pass of Fractionation.

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 $n_g \leq n_c$ 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.

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 [11] 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. Let $n_{i\cdot}$ be the sum over the i -th row of the table, and let $n_{\cdot j}$ be the sum over the j -th column. Then the Fowlkes-Mallows index is given by:

$$\sum_{i,g} \binom{n_{ig}}{2} / \sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_g \binom{n_{\cdot g}}{2}} \quad (4)$$

true groups	clusters				Total
	1	2	...	G	
1	n_{11}	n_{12}	\cdots	n_{1G}	$n_{1.}$
2	n_{21}	n_{22}	\cdots	n_{2G}	$n_{2.}$
...	\cdots	\cdots	\cdots	\cdots	\cdots
J	n_{J1}	n_{J2}	\cdots	n_{JG}	$n_{J.}$
Total	$n_{.1}$	$n_{.2}$	\cdots	$n_{.G}$	n

Table 2: Comparison between clusters (columns) and groups (rows). Each cell count n_{ig} is the number of common elements in cluster g and group i .

4.2 The TDT dataset

Our examples are derived from a dataset of 1,131 documents that is part of the Topic Detection and Tracking document collection [1]. The 1,131 documents were manually classified into a total of 25 groups or topics. Six of the topics have less than eight documents each, and contain a total of only 31 documents. We only used the remaining 1,100 documents, partitioned into 19 topics.

We relied on standard document retrieval technology to convert the 1,100 documents into vectors in a 50-dimensional space: We assembled the term-document matrix, applied the log-Idf transformation to the term counts as suggested by Dumais [9], and then reduced the dimensionality by latent semantic indexing [4].

Model-based clustering of the 1,100 documents (more precisely, the 1,100 50-dimensional vectors corresponding to the documents) into 19 clusters resulted in a Fowlkes-Mallows index of 0.76 and a 19×19 contingency table (analogous to Table 2) with 27 non zero entries. This is the standard against which we measure the results in the following examples.

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 $M = 1,000$ 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 AWE criterion. We repeated this experiment — generating a sample of size 22,000 and clustering it — ten times. In these ten replications, the AWE criterion 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 Figure 9), indicating almost perfect

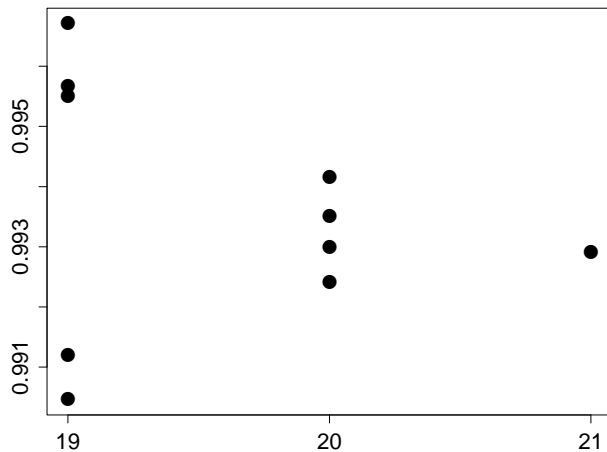


Figure 9: Fowlkes-Mallows index vs number of clusters chosen by AWE for Example 1.

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 [17] 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. 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 AWE criterion for choosing the final number of clusters. In ten replications of the experiment the AWE criterion 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 Figure 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.

We generated the data for Example 3 by essentially replicating the TDT dataset 19 times, replacing each group by a scaled and shifted version of the entire dataset: Let μ_i and Σ_i be the mean vector and covariance matrix of the i -th group. We obtained the i -th replicate by scaling and shifting the entire dataset to have mean vector μ_i and covariance matrix Σ_i . We end up with

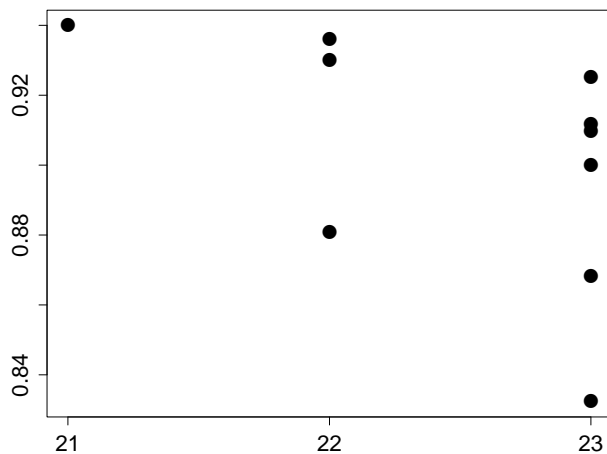


Figure 10: Fowlkes-Mallows index vs number of clusters chosen by AWE for Example 2.

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

Table 3: Example 3 – agreement between clusters and groups after each Fractionation pass.

$19 \times 19 = 361$ groups and $19 \times 1100 = 20,900$ observations.

We randomly split these 20,900 observations into $M = 20$ fractions of 1,045 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

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

Table 4: Example 3 – distribution of the number of fractions in which groups are represented, 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

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

split across two fractions.

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 AWE criterion, 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 [14, 15], combined with a criterion like AWE 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.

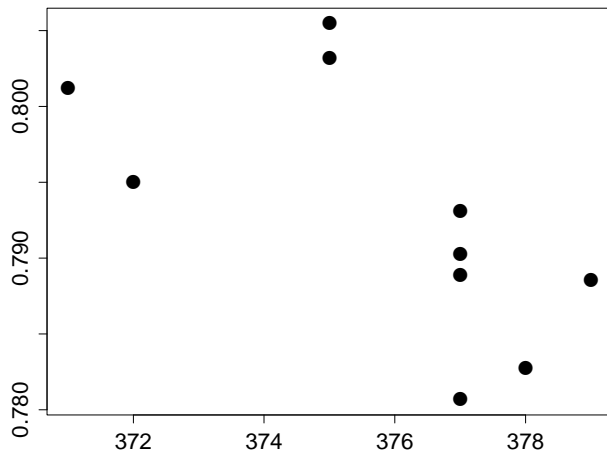


Figure 11: Fowlkes-Mallows index vs number of clusters chosen by AWE for Example 4.

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 ten 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 AWE criterion. In ten replications of this experiment, the number of clusters chosen by the AWE criterion was between 371 and 379, with values of the Fowlkes-Mallows index between 0.781 and 0.806 (see Figure 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 AWE criterion.

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 AWE criterion 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] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report, 1998.
- [2] M. Ankerst, M. Breuning, H. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings, ACM SIGMOD International Conference on Management of Data (SIGMOD'99)*, pages 49–60, 1999.
- [3] J. D. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.
- [4] M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.
- [5] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large datasets. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD98)*, 1998.
- [6] P. Bradley, U. Fayyad, and C. Reina. Scaling EM (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35, Microsoft Research, 1999.
- [7] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *15th Ann Int'l SIGR*, pages 318–329, 1992.
- [8] P. Domingos and G. Hulten. Learning from infinite data in finite time. In *Advances in Neural Information Processing Systems 14*. 2002.
- [9] S. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments & Computers*, 23(2):229–236, 1991.
- [10] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, 1996.
- [11] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *J. American Statistical Association*, 78:553–569, 1983.

- [12] C. Fraley and A. Raftery. How many clusters? which clustering method? - answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.
- [13] J. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.
- [14] G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8(1):65–74, 1997.
- [15] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.
- [16] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:497–511, 1978.
- [17] D. Scott. *Multivariate Density Estimation*. Wiley, 1992.
- [18] D. Wishart. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In A. Cole, editor, *Numerical Taxonomy*, pages 282–311. Academic Press, 1969.