Handling Outliers through Agglomerative Clustering with Full Model Maximum Likelihood Estimation, with Application to Flow Cytometry

Mark Gordon, Justin Li, Kevin Matzen, Bryce Wiedenbeck

Motivation

Clustering is an important and useful process in many data processing fields. In many of these fields, the base distribution of the data is unknown. Where partitional clustering algorithms such as Gaussian mixture models are used, the incorrect model assumption causes a significant portion of the data to appear as outliers. This problem is compounded by the fact that mixture models, which are based on maximum likelihood estimation, are sensitive to the presence of outliers. To obtain a better clustering performance, an alternative is to use hierarchical clustering algorithms, which create a tree of nested clusters based on the dissimilarity between points. Intuitively, outliers are the least similar to other points in the data, and therefore will be easily identified.

One motivating example where it is particularly important to reduce the sensitivity of clustering algorithms to outliers is flow cytometry. In flow cytometry, the different dimensions of data points correspond to the florescence response of different wavelengths of cells. Data from this field tend to form rough clusters related to cell type, but outliers are common due to dead cells and other cell debris. Traditionally, flow cytometry data are clustered manually by an expert operator, by thresholding the high dimensional data in 2 and 3 dimensional projections. An automated approach to this clustering problem that was robust to outliers could substantially improve the speed and accuracy of this process.

Technical Background

Agglomerative clustering is a family of clustering methods where each data point is initially assigned to its own singleton cluster. At each step of the algorithm, the two most similar (least dissimilar) clusters are merged until every data point belongs to a single cluster. This repeated merging can be represented by a tree-like structure, called a dendrogram, which captures the order of the merging as well as the dissimilarity between the merged cluster.

The core of any agglomerative clustering algorithm is the dissimilarity measure used to select which clusters to merge. One traditional measure is Ward's method, where the dissimilarity is defined to be the difference between the error sum of squares (ESS) of the merged cluster and the ESS of the two original clusters. The mathematical definition of Ward's method is shown below, where C_1, C_2 are the clusters to be merged:

$$d_{ward}(C_1,C_2) = ESS(C_1 \cup C_2) - ESS(C_1) - ESS(C_2)$$

where a cluster C with mean μ has the error sum of squares:

$$ESS(C) = \sum_{x \in C} (x - \mu)^2$$

(Kamvar et al. 2002) demonstrated model-based interpretations for Ward's method and other classical dissimilarity measures, where each merge is a greedy step in maximizing the likelihood of fitting the underlying model. Ward's method is equivalent to greedily maximizing the likelihood at each step assuming an equal-variance isotropic Gaussians as the underlying distribution. This suggests that principled extensions to agglomerative clustering may be developed by using a different underlying model to create new dissimilarity metrics.

It is important to note that, although extensions of this form also make distributional assumptions, they remain insensitive to outliers. The model assumption is used to decide which clusters to merge, not the final cluster assignment. As such, singleton clusters which are merged late in the algorithm can be identified as outliers. This extension is also different from running a mixture model expectation maximization (EM) algorithm repeated over the data with different numbers of components, as EM clusters from one run have no particular structural relation to the clusters from another run with a different number of components (Zhong et al. 2003).

Innovation

To show equivalence between a dissimilarity measure and a model assumption, define a likelihood function of the model producing a partition with *k* clusters $C_{1,...,}C_k$. This is equivalent to the likelihood of a *k*-component mixture model. Implicitly, the parameters for each component θ_i are the ones which maximize the likelihood of producing the data in that component. The likelihood, J, of a partition $P_k = \{C_{1,...,}C_k\}$ is therefore:

$$J(P_k) = \prod_{i \in \{1,...,k\}} \max_{\{\theta_i\}} L(\theta_i; \{x_j \in C_i\})$$

The cost of a merge from P_k to P_{k-1} is

$$\Delta J(P_k, P_{k-1}) = \frac{J(P_{k-1})}{J(P_k)}$$

To maximize the likelihood of a partition, at each step the agglomerative clustering algorithm selects the partition P_{k-1} which maximizes the above. Since the algorithm is written to merge the two least dissimilar clusters, it is natural to create dissimilarity measures which are negative multiples of the cost:

$$d(C_i,C_j) = -c\Delta J(\{C_i,C_j\},\{C_i \cup C_j\}) = -c\frac{J(\{C_i \cup C_j\}}{J(\{C_i,C_j\})}$$

Although the classical methods are model based, they do not in fact use the full model; Ward's method, which assumes equal-variance isotropic Gaussians, reduces the likelihood estimation to the ESS calculation above. This can be easily improved by using the full multivariate Gaussian distribution. That is, take the likelihood function for a particular cluster C_i to be:

$$L_N(\theta_i;\{x_j \in C_i\}) = L(\mu_i,\Sigma_i;\{x_j \in C_i\}) = \prod_{x_j \in C_i} \phi(x_j;\mu_i,\Sigma_i)$$

In a similar vein, other distributions can be substituted into an agglomerative
Justin did the significant majority of the work on experiment 1, Bryce was responsible for experiment clustering algorithm. Noting the use of skew-normal distributions in flow cytometry software (Hu 2009), the following likelihood function is defined:

$$L_{SN}(\theta_i;\{x_j \in C_i\}) = L(\mu_i,\sigma_i,\alpha_i;\{x_j \in C_i\}) = \prod_{x_j \in C_i} f(x_j,\mu_i,\sigma_i,\alpha_i)$$

where

$$f(x;\mu,\sigma,\alpha) = 2\phi(x-\mu;\Sigma)\Phi(\alpha^T w^{-1}(x-\mu))$$

is the probability density function of the skew-normal distribution (Azzalini and Capatanio 1998), with α being the skewness.

For simplicity, the two dissimilarity metrics above will be referred to as NMLE link (for normal distribution maximum likelihood estimation) and SNMLE link (for skew normal MLE).

Agglomerative clustering is inherently less susceptible than other clustering methods to distortion by outliers because outliers are necessarily the least similar to other data points. However, as clusters grow, the model likelihood of combining two large clusters will approach that of combining one large cluster with its nearest outliers. At this point, it becomes important to detect outliers, as they will affect further likelihood estimates after they have been merged with a major cluster. This detection can be accomplished by comparison of relative cluster size at each merge. If one cluster is drastically larger than the other, it is likely that the small cluster contains outliers.

Methodology

Two main experiments were conducted. The first is a comparison between the full model dissimilarity measures above and other clustering methods, including K-means, a Gaussian mixture EM algorithm, and agglomerative cluster with Ward's method. The second experiment explores one heuristic for explicit outlier detection in agglomerative clustering.

In the first experiment, synthetic data was generated from a 2 dimensional mixture model with 2 components. Two different distributions were used - the multivariate normal distribution, as well as the multivariate skew-t distribution, which has been applied with success to flow cytometry data (Pyne et al. 2009). In addition to varying the underlying distribution, uniform noise was added. The signal to noise ratio varied over the range from 0% to 20% noise. The total number of points was varied for the NMLE link experiments, to see how sparse data affected the metric. The results are the average of 100 trials with each combination of parameters. The graphs below show the two distributional mixture models with and without noise:

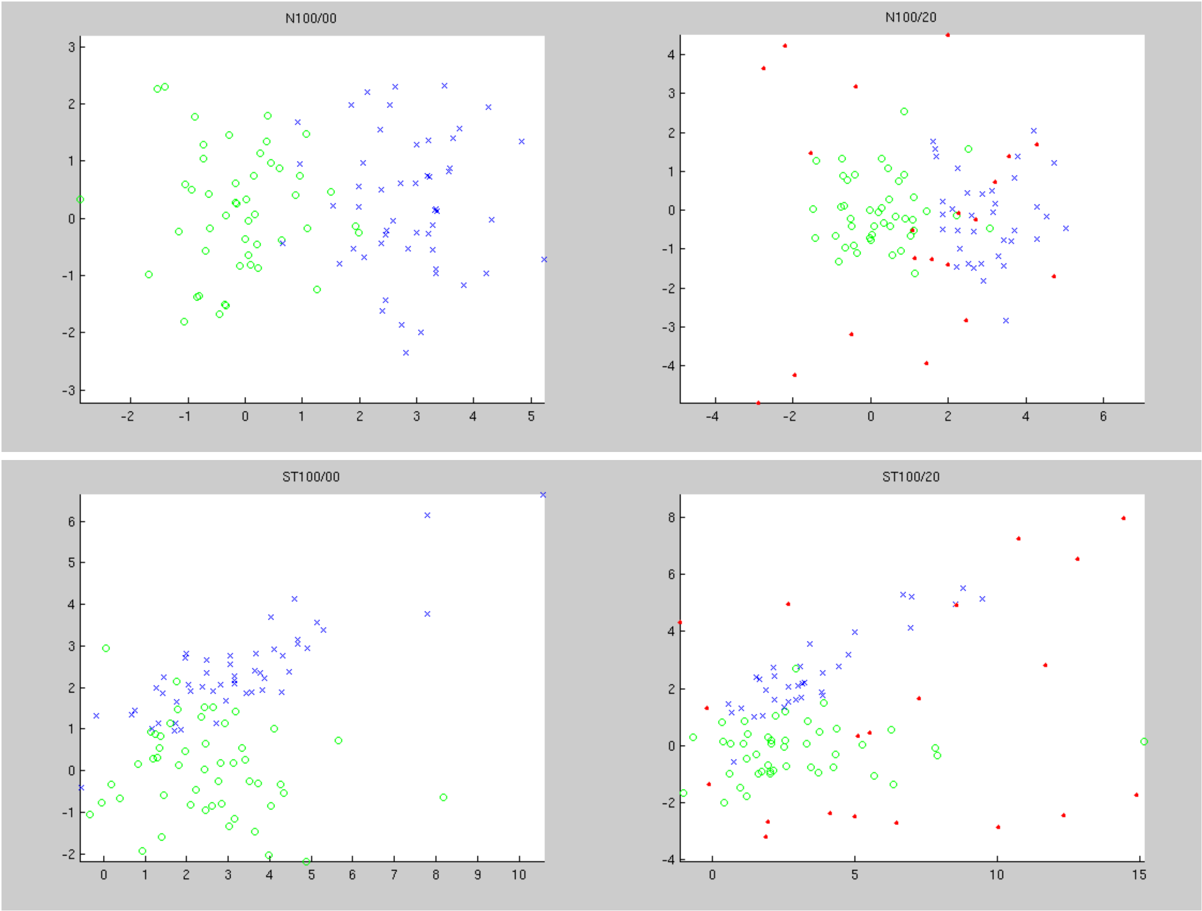


Figure 1: synthetic data to test clustering algorithms. The top two graphs are drawn from a 2 component Gaussian mixture, while the other two are from a 2 component skew-t mixture. The right graphs are noisy (with 20% of points being uniform noise).

For evaluation, the Rand index was used to compare the label partition and the result of the clustering algorithm. Rand index for two partitions P and Q both with n objects is calculated as follows (Hubert and Arabie 1985):

$$RI(P,Q)=\frac{a+b}{\binom{n}{k}}$$

where *a* is the number of pairs of objects where both objects have the same label in *P* and both have the same label in *Q*, and *b* is the number of pairs of objects where both objects have different labels in *P* and different labels in *Q*. A Rand index of 0 indicates the no two objects are in the same cluster in both partitions, while a Rand index of 1 indicates that all pairs of objects are either in the same cluster in both partitions, or are separate in both partitions.

The second experiment explored a method for automatic outlier detection within the framework of agglomerative clustering. The experiment used data from a Gaussian mixture model plus fixed-ratio (10%) random noise, and the agglomerative algorithm was modified to check the ratio of the sizes of each pair of clusters being merged. If the ratio was above a fixed threshold, the merge was rejected and the smaller cluster marked as outliers. This heuristic could potentially augment any of the model-based methods examined in experiment one.

Experimental Results - Comparison of Clustering Algorithms

The result of the first experiment is presented in Table 1. The parameters field is labeled first by underlying distribution (N for normal, ST for skew-t), then by the number of points (50 or 100), and finally by percentage noise (0%-20%). See Figure 1 for sample draws

Parameters	K-Means	GMM EM	Ward's Method	NMLE Link	SNMLE Link	Best
N50/00	0.874	0.739	0.837	0.644	0.529	K-Means
N50/05	0.865	0.701	0.825	0.613	0.524	K-Means
N50/10	0.859	0.661	0.825	0.629	0.531	K-Means
N50/15	0.852	0.629	0.819	0.604	0.521	K-Means
N50/20	0.772	0.610	0.795	0.586	0.504	Ward's
N100/00	0.877	0.781	0.827	0.656	-	K-Means
N100/05	0.869	0.669	0.821	0.653	-	K-Means
N100/10	0.856	0.648	0.811	0.641	-	K-Means
N100/15	0.843	0.591	0.795	0.636	-	K-Means
N100/20	0.842	0.580	0.815	0.620	-	K-Means
ST50/00	0.541	0.611	0.547	0.566	0.510	GMM
ST50/05	0.530	0.556	0.523	0.557	0.519	NMLE
ST50/10	0.516	0.530	0.515	0.530	0.520	NMLE

ST50/15	0.511	0.519	0.509	0.538	0.509	NMLE
ST50/20	0.511	0.523	0.518	0.529	0.508	NMLE
ST100/00	0.537	0.600	0.536	0.571	-	GMM
ST100/05	0.521	0.534	0.529	0.548	-	NMLE
ST100/10	0.511	0.527	0.520	0.539	-	NMLE
ST100/15	0.514	0.510	0.521	0.525	-	NMLE
ST100/20	0.506	0.512	0.510	0.525	-	NMLE

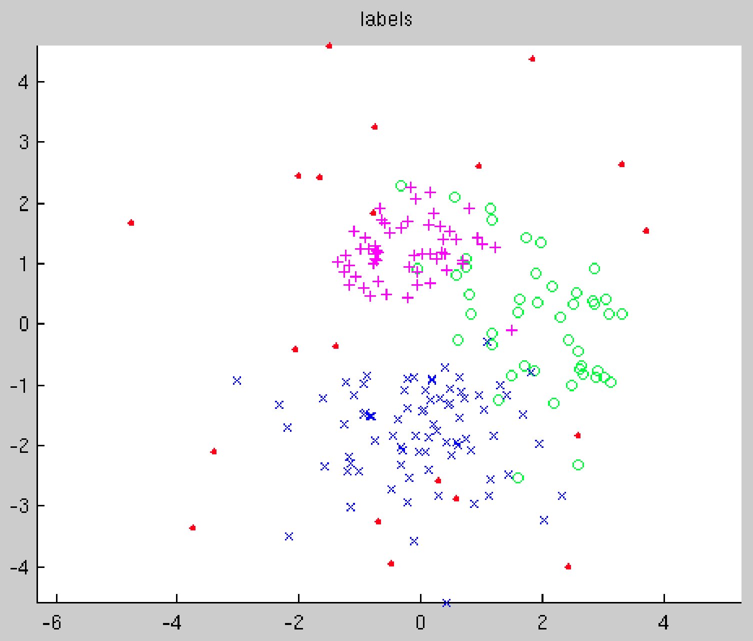
Table 1: Experimental results from testing 5 different clustering algorithms. The metric used is the Rand index; 0 indicates the algorithm resulted in entirely different clusters, while 1 indicates the algorithm was perfect.

As the above table shows, full model MLE linkages work better than the other methods when the model assumptions are incorrect. Although the SNMLE linkage metric in theory contains the NMLE linkage (by setting the skewness to 0), in practice it performs worse due to a greater number of parameters to estimate.

Experimental Results - Outlier Detection

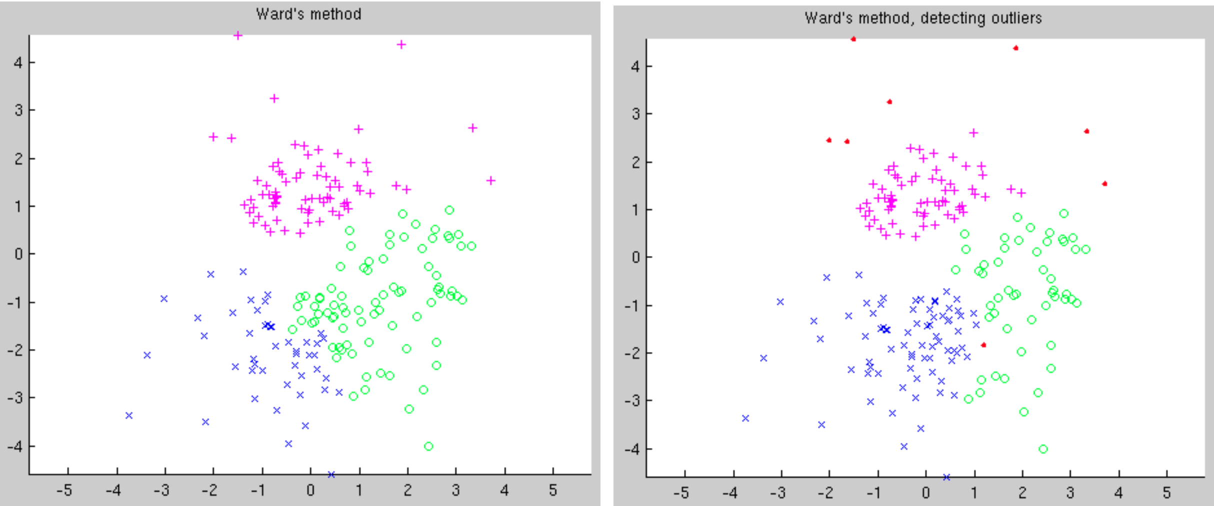
The second experiment is presented in Figure 2. 180 data points were generated by a mixture of three Gaussians, and 20 by a uniform random distribution over the figure area: labeled data are shown in (a). (b) gives results from agglomerative clustering using Ward's method dissimilarity measure, with and without detection of outliers. K-means and EM were also run on this data set, resulting in Rand indices of 0.81 and 0.555, respectively.

Figure 2:



(a) Labeled data:

red dots ~ uniform over $(-5,5)\times(-5,5)$ blue x's ~ $G(\mu=(0,-1);\sigma=[1,0;0,1])$
green circles ~ $G(\mu=(0,2);\sigma=0)$ pink pluses ~ $G(\mu=(-1,1);\sigma=0)$

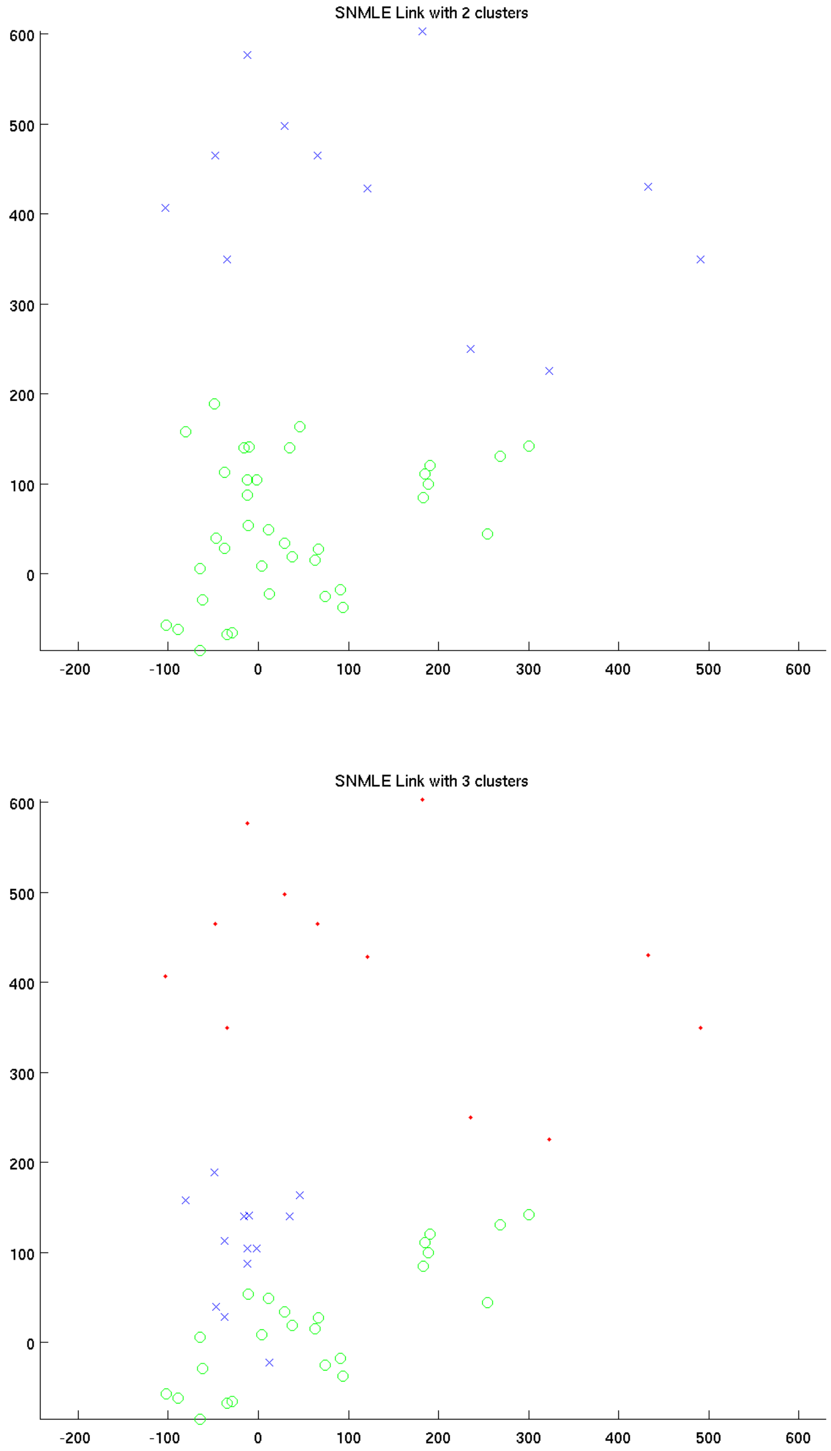


(b) Left: Ward's method, Rand index = 0.6432
Right: outlier detection, Rand index = 0.8168

As the Rand index values demonstrate, augmenting agglomerative clustering with active outlier detection can further improve the algorithm's robustness.

Experimental Results - Flow Cytometry Data

Combining the two approaches above, SNMLE link with outlier detection was applied on a portion of the flow cytometry data. The sample size is small (n=50) due to memory and run time issues; however, it demonstrates that SNMLE link finds plausible sub-populations within the data. However, due to the computational resources required to use agglomerative clustring with SNMLE link much of the method seems impractical as normally much larger data sets would be used.



SNMLE Link with 2 clusters

SNMLE Link with 3 clusters

Figure 3: Applying SNMLE link to a portion of flow cytometry data (n=50). Note that denser sub populations are found.

From the results of the experiments it would seem useful to calculate clustering using this method however it does not seem practical. One potential avenue for future research would be to see if an equivalent dissimilarity metric could be computed more efficiently in a manner similar to how Ward's dissimilarity metric is a more efficient equivalent to an Isotropic Gaussian MLE link.

Discussion

The comparison of different clustering algorithms demonstrates that full model MLE linkages perform better than other clustering methods when the underlying model assumptions are incorrect. Compared to the Gaussian distribution, the t distribution has more probability mass on the tails. Together with the fact that the distribution is skewed, a significant number of points are sampled away from the non-tail "body". For a GMM EM algorithm, these points shift the distribution towards the tail, which can cause the algorithm to incorrectly cluster the central points.

In agglomerative methods, on the other hand, these outliers have less of a direct effect on the labels given to the central points. The greater density of points in the body increases the likelihood of those points being drawn from the same distribution, which is what the greedy step in t he algorithm tries to maximize. Notably, this likelihood has no relation to the tail; whereas an EM algorithm maximizes the global likelihood, the likelihood estimation is local in an agglomerative clustering algorithm. This serves to clearly demarcate the dense clusters before any outliers are considered.

There are, however, distinct disadvantages to agglomerative clustering. Especially in flow cytometry where the sample is large, the $O(n^2)$ running time of creating dissimilarity matrices is undesirable. Together with the fact that MLE is also an iterative algorithm, this makes full model MLE linkages have an extremely slow running time.

Conclusion

Agglomerative clustering is a useful technique for unlabeled data that do not fit clearly to a particular mixture model. The local nature of agglomerative merges makes the base algorithm relatively robust to data outliers. In addition, model-based approaches allow agglomerative clustering to fit a wide variety of data types. Finally, explicit outlier detection can further reduce the impact of outliers on correct clustering of the majority of data points.

Despite the robustness of agglomerative clustering with MLEs, its computational complexity is a significant drawback, which requires careful consideration if it is applied to large, high-dimensional data sets.

References

A. Azzalini and A. Capitanio (1998). Statistical applications of the multivariate skew—normal distribution. *Journal of the Royal Statistical Society* , series B vol. 61 pp. 579—602.

L. Hubert and P. Arabie (1985). Comparing Partitions. *Journal of Classification* , vol. 2 pp. 193—218.

R Castro, M. Coates and R. Nowark (2004). Likelihood based hierarchical cluster. In *IEEE Transactions in Signal Processing*, vol. 52 no. 8 pp. 2308—2321. .

S. Kamvar, D. Klein and C. Manning (2002). Interpreting and extending classical agglomerative clustering algorithms using a model—based approach. In *Proceedings of the 19th Conference on Machine Learning*, 283—290. .

S. Pyne, B. What and C. What (2009). Automated high—dimensional flow cytometric data analysis. In *Proceedings of the National Academy of Sciences*, vol. 106 issue 21 pp. 8519—8524. .

S. Zhong and J. Ghosh (2003). A unified framework for model—based clustering. *Journal of Machine Learning Research* , vol. 4 pp. 1001—1037.

X. Hu (2009). *FLAMEMixtureModel Documentation*. Broad Institute.

Contributions

Justin did the significant majority of the work on experiment 1, Bryce was responsible for experiment 2, and Mark worked on the flow cytometry application. Justin wrote much of the report, with help from Bryce. Bryce and Mark edited the report draft. Kevin contributed in the early stages of the project.