# Uncovering Group Level Insights with Accordant Clustering: Supplementary Material

Amit Dhurandhar* Margareta Ackerman† Xiang Wang‡

## Appendix

### Proof of Theorem 1

*Proof.* We would like to find the probability that each center lies in a distinct cluster core. Consider the event in which one of the cluster cores (say the core of cluster $C_i$) is missing from the selected $k$ centers. The probability of this event occurring is:

$$p_i \leq \left(1 - \frac{|C_i^o|}{|X|}\right)^k \leq (1-\epsilon')^k \leq e^{-\epsilon' k}.$$

Let $\theta$ be the probability this event *does not* occur for any of the $k$ clusters - that is, there is a center in each cluster core. Then,

$$\theta \geq 1 - \sum_{i=1}^{k} p_i \geq 1 - ke^{-\epsilon' k}.$$

Now let's assume that we successfully pick a center in each cluster core. Then we assign each point to its closest center, which leads to an $(r,t)$-accordant clustering $C'$. By the proof of Lemma 2, the algorithm changes this clustering as it continues to iterate towards convergence only if it can find a lower cost solution.

The optimal $(r,t)$-accordant clustering $C_A$ can only have cost better than $C'$. Since $C'$ has near-optimal cost, so does $C_A$ (the optimal $(r,t)$-accordant clustering). As our algorithm outputs a clustering that has cost no higher than that of $C'$, it outputs a clustering of near-optimal cost. By the $(\alpha,\epsilon)$-property the clustering it finds is $\epsilon$-close to the optimal clustering $C^*$, which is $\epsilon$-close to the optimal $(r,t)$-accordant clustering $C_A$. As such, we find a clustering that is $2\epsilon$-away from the optimal accordant solution.

### Proof of Corrolary 1

*Proof.* The probability that one of the cluster cores is missing from a selection of $k$ centers is $p_i \leq e^{-\epsilon' k}$, as shown in Theorem 1. If $\theta$ is the probability that this *does not* occur for any of the $k$ clusters *at least once* in $m$ trials, then:

$$\theta \geq 1 - (\sum_{i=1}^{k} p_i)^m \geq 1 - (ke^{-\epsilon' k})^m$$

For any iteration with a center in each core, by the proof of Theorem 1, the algorithm will find a clustering that is $\epsilon$-close to the optimal clustering $C^*$ and $2\epsilon$-close to the optimal $(r,t)$-accordant solution.

Once a near-optimal clustering is found, any subsequent clusterings will only be chosen if they have lower cost. Hence, our algorithm outputs an $(r,t)$-accordant clustering of near-optimal cost that is $2\epsilon$-close to the optimal $(r,t)$-accordant solution.

**UCI datasets** We now evaluate the methods on 6 UCI datasets used in previous clustering studies [1] namely: a) Glass, b) Heart, c) Ionosphere, d) Breast Cancer, e) Iris and f) Wine. The Glass dataset, the Heart dataset, the Ionosphere dataset, the Breast Cancer dataset, the Iris dataset and the Wine dataset are partitioned into 6 groups, 2 groups, 2 groups, 2 groups, 3 groups and 3 groups respectively, as indicated by their ground truth labels.

The performance of the various methods on these datasets is seen in figure 1. $k$ is set to the number of groups and we vary $t$ in each case with $r$ being set to 1. Given the space constraints, we plot the mean and the confidence intervals in the figures themselves. However, to keep the exposition clear we depict the results using bar charts for low, medium and high values of $t$. In particular, we compare the different methods for $t = \{0.2, 0.5, 0.8\}$.

We see from the figure that across all the datasets Akmeans matches the performance of $k$-means for low and medium values of $t$ when $k$-means satisfies our constraint. This again reaffirms the fact that our method can provide the same quality clustering as standard $k$-means when our constraint is trivially satisfied. The other methods have consistently higher (mean) error and in some cases even higher variance than our method.

For high values of $t$, where $k$-means does not satisfy our constraint such as on Heart, Ionosphere and Wine,

*IBM Research, adhuran@us.ibm.com
†San Jose State University, margareta.ackerman@sjsu.edu
‡Google, xiangwa@google.com

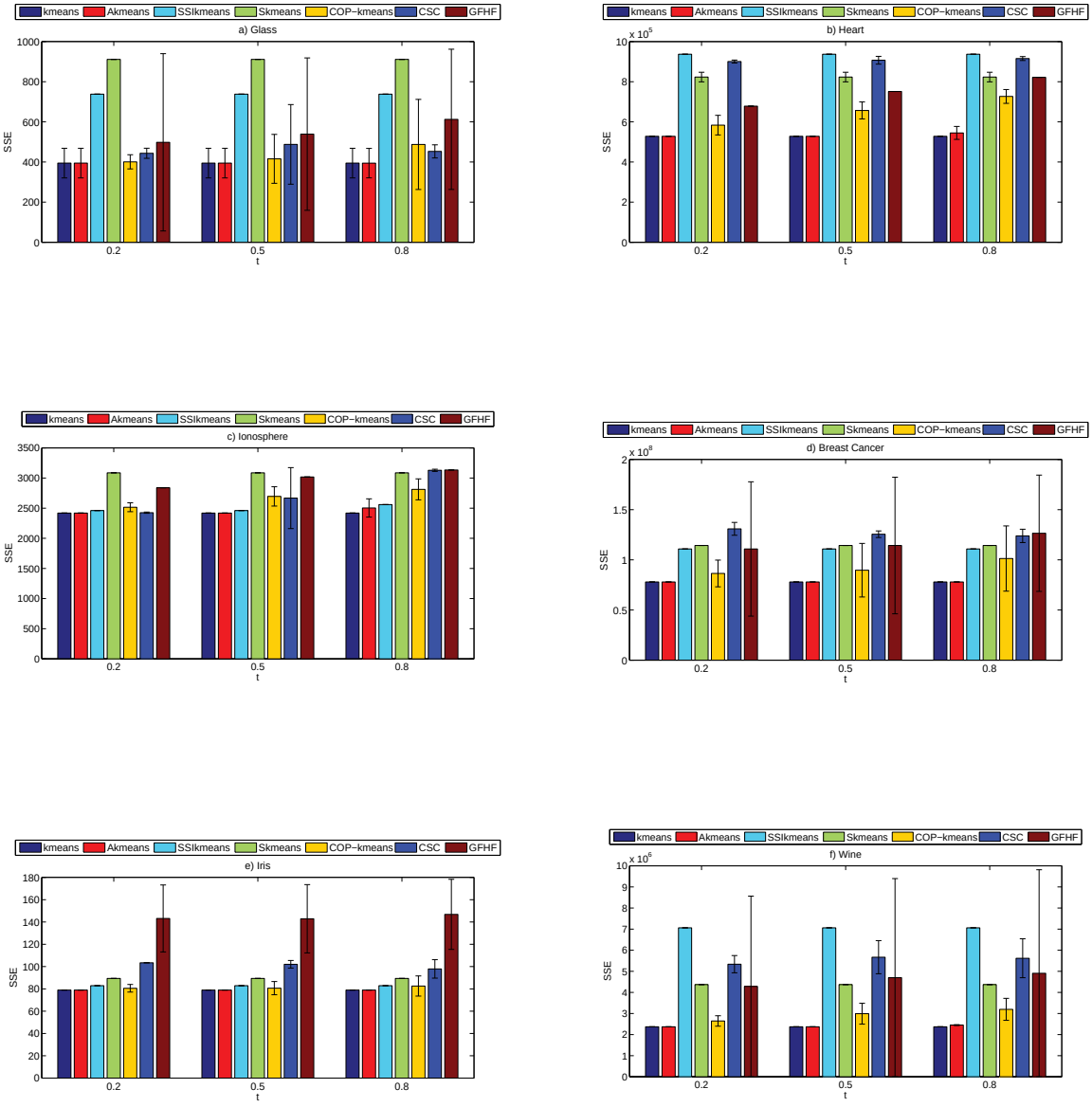


Figure 1: Above we see the performance (mean +- 95% confidence interval) of the various methods on 6 UCI datasets namely: a) Glass, b) Heart, c) Ionosphere, d) Breast Cancer, e) Iris and f) Wine for 3 different (low, medium, high) values of $t$. The bars for which we do not see any confidence interval correspond to runs that have zero or insignificant variance.

| $k$ | kmeans | Akmeans | Skmeans | SSIkmeans | COPkmeans | CSC | GFHF |
|---|---|---|---|---|---|---|---|
| 2 | 41.2 | 41.2 | 88.3 | 749.9 | 2519.4 | 1899.2 | 3212.4 |
| 3 | 31.5 | 21.4 | 183.5 | 698.8 | 2245.3 | 1834.2 | 39344.9 |
| 4 | 9.5 | 22.5 | 456.3 | 677.7 | 2698.1 | 1829.5 | 4171.9 |
| 5 | 10.2 | 21.3 | 491.1 | 594.6 | 2746.8 | 1704.4 | 42840.1 |
| 6 | 13.5 | 57.6 | 258.5 | 619.6 | 2734.4 | 1840.4 | 4690.4 |

Table 1: The above table shows half the width of the 95% confidence interval (based on the randomizations) for the different methods and for different values of $k$ around the corresponding means w.r.t. the Health Care dataset.

| $k$ | kmeans | Akmeans | Skmeans | SSIkmeans | COPkmeans | CSC | GFHF |
|---|---|---|---|---|---|---|---|
| 2 | 513.5 | 513.5 | 482.3 | 6492.5 | 22199.8 | 16595.7 | 37903.4 |
| 3 | 297.2 | 1999.2 | 1571.5 | 5973.5 | 19225.3 | 16112.2 | 45340.9 |
| 4 | 117.8 | 2244.6 | 3867.8 | 6173.8 | 23943.1 | 17581.5 | 48660.2 |
| 5 | 100.2 | 2148.9 | 4117.1 | 5292.6 | 25362.8 | 26589 | 50149.1 |
| 6 | 520.5 | 1574.9 | 2470.3 | 5382.3 | 25700 | 16742 | 48677.4 |
| 7 | 420.7 | 2177.1 | 3756.7 | 4836.2 | 25234.7 | 11883.6 | 50216.9 |

Table 2: The above table shows half the width of the 95% confidence interval (based on the randomizations) for the different methods and for different values of $k$ around the corresponding means w.r.t. the Spend dataset.

Akmeans is only incrementally worse than $k$-means, though it provides a feasible clustering. In this case too, when $k$-means does provide a feasible clustering our method outputs the same quality clustering. The other methods are much worse in most cases with again higher error and in some cases higher variance.

Amongst the other methods COPkmeans seems to be performing the best overall with lower error and moderate variance in many cases. However, its higher error and in many cases higher variance relative to our method is again because of its sensitivity to the chosen $t$ fraction that it must assign to the same cluster. This gap is much lesser on the Iris dataset, since the groups are relatively well separated with not much overlap. The sensitivity issue is also prevalent in CSC for the same reasons. For GFHF besides the sensitivity to the $t$ fraction its performance is also affected by the instances we choose from the other groups to initialize it. This is the reason for its excessively high variance in multiple cases.

The supervised methods perform much worse than our method in general, since they strive to cluster all instances in a manner that is consistent with the groups they belong to. Hence, this procedure turns out to be excessively demanding leading to lower quality feasible clusterings.

**References**

[1] X. Wang, B. Qian, and I. Davidson. Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering. In *ICDM*, pages 1146–1151. IEEE Computer Society, 2012.