Lightweight Clustering Methods on Population SNP Data

Isaac Thomas

March 15, 2024

0 Introduction

0.1 Motivation

Single Nucleotide Polymorphisms (SNPs) have had massive intergenerational effects on human phenotypes. Such impact renders SNP data useful in describing phenotypes to an extent that allows informative clustering. As biologically/probabilistically informed methods for this task can be more space/computation intensive, we attempt to determine the simplest clustering method that can informatively cluster SNP data with the goal of capturing population demographics. Using the 1000 genomes dataset, we compare k-means and hierarchical clustering regarding how much structure they can discern without substantial biological or probabilistic information/assumptions. We also highlight limitations of this work and avenues for future research.

1 Materials & Methods (Available Here)

1.1 Data & Preprocessing

We used the 1000 Genomes Dataset [1], which consists of SNPs from thousands of individuals categorized by population. To save time, we used only SNPs from chromosome 19. We used data that was already LD-pruned to eliminate SNPs strongly correlated for reasons unrelated to population traits. We then performed principal component analysis with `plink`, keeping 4 principal components. Given our SNP data matrix X, PCA computes the singular value decomposition $\boldsymbol{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where the rows of the lower-dimensional $\mathbf{U}$ are the transformed data in the basis determined by the columns of $\mathbf{V}^\top$. The columns of $\mathbf{V}^\top$ contain the contribution of each SNP to the data.

1.2 Clustering Methods Used

We first used k-means clustering with Lloyd's algorithm [3] for iterative centroid computation and point-cluster reassignment. For this experiment and the following one, we used silhouette score [5] as a measurement of cluster quality, which measures the "tightness" of the inferred clusters. We tuned k to maximize silhouette score (500 iterations per clustering) using `scikit-learn` [4].
For hierarchical clustering, we used an *agglomerative* approach which treats each point as a cluster and repeatedly merges the two "closest" clusters until some stopping criterion is met. We determined the closest clusters at each iteration via the following methods for two clusters A and B:

- single linkage [2]: take the minimum distance between any point in A and any point in B.
- complete linkage [2]: take the maximum distance between any point in A and any point in B.
- average linkage [6]: take the average distance between any point in A and any point in B.
- Ward linkage [7]: take the resulting change in intra-cluster variance by merging A and B.

Using `scikit-learn`, we tuned all hierarchical approaches by distance threshold, which is the distance between two clusters that renders them too far away to merge (stopping criterion).

2 Discussion

2.1 Results

We found that k-means clustering and hierarchical clustering performed similarly in extracting insight from the PCA-transformed SNP data. Through hyperparameter tuning w.r.t. silhouette score, k-means discerned an optimal 5 clusters, which was near the correct number of populations in the data (6). Figure 1 shows the resulting visualization, with clear separation between most of the big populations.
k-means attained a silhouette score of ~ 0.75 using the optimal value of k. A score near 0 would indicate lots of overlapping clusters, and a score below 0 would indicate a very large number of misassignments. As silhouette score lies in $[-1, 1]$, this indicates that the clustering is decently tight. Although there is decent separation between the populations, there is little separation within them. This may foreshadow difficulty in discerning subpopulation structure, which turned out to be the case as shown in figure 2.

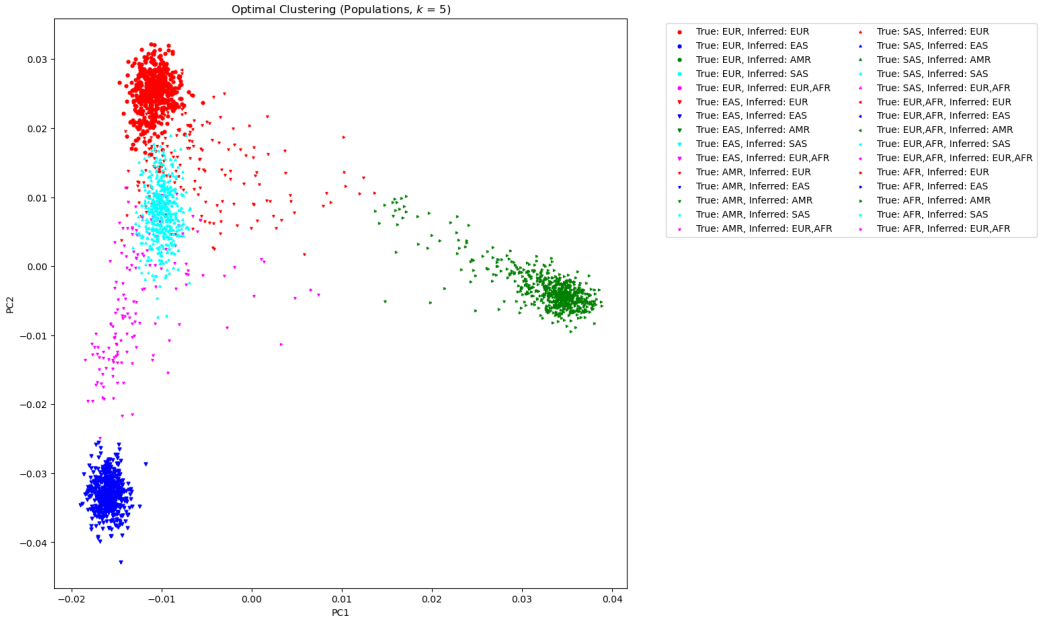


Figure 1: Results of tuning k in the clustering to something consistent with the number of subpopulations. The shape of each point is its true subpopulation, and its color is the subpopulation it was assigned to.

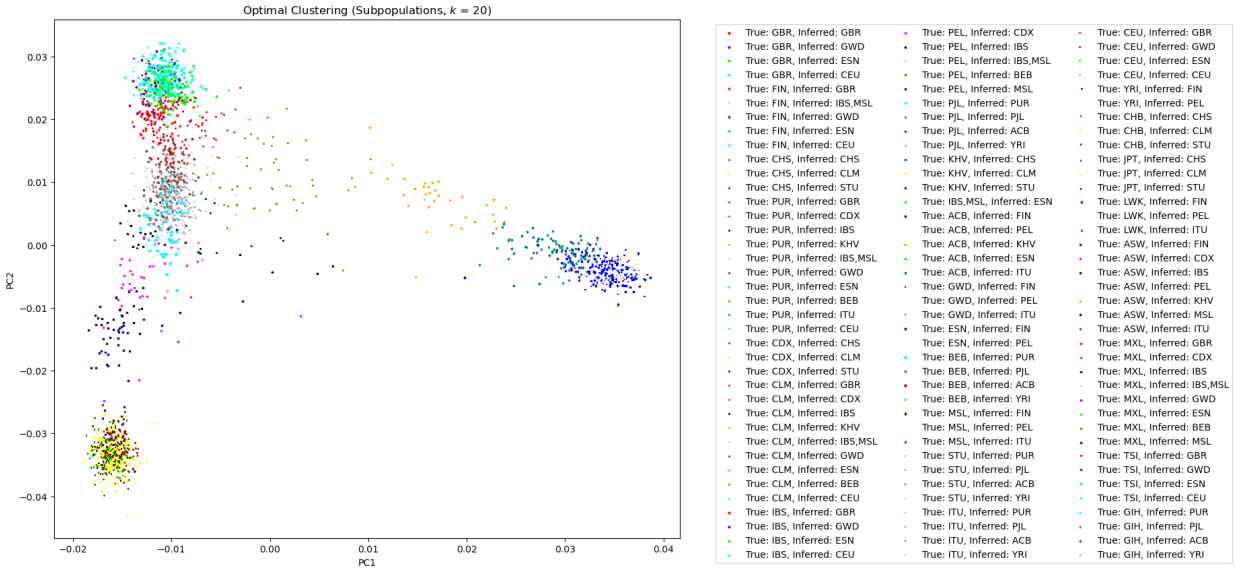


Figure 2: Results of k-means clustering on the the LD-pruned SNP 1000 Genomes data (Chr 19). The shape of each point is its true subpopulation, and the color indicates the subpopulation inferred by k-means.

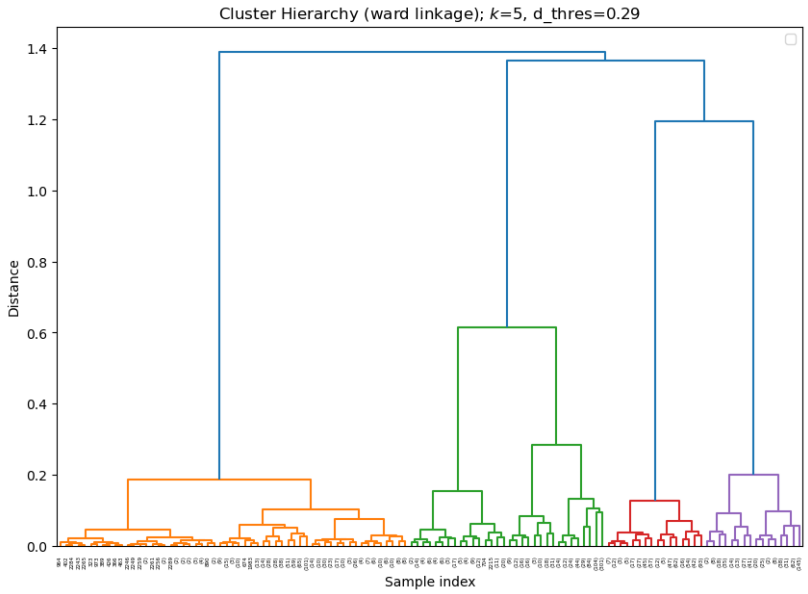


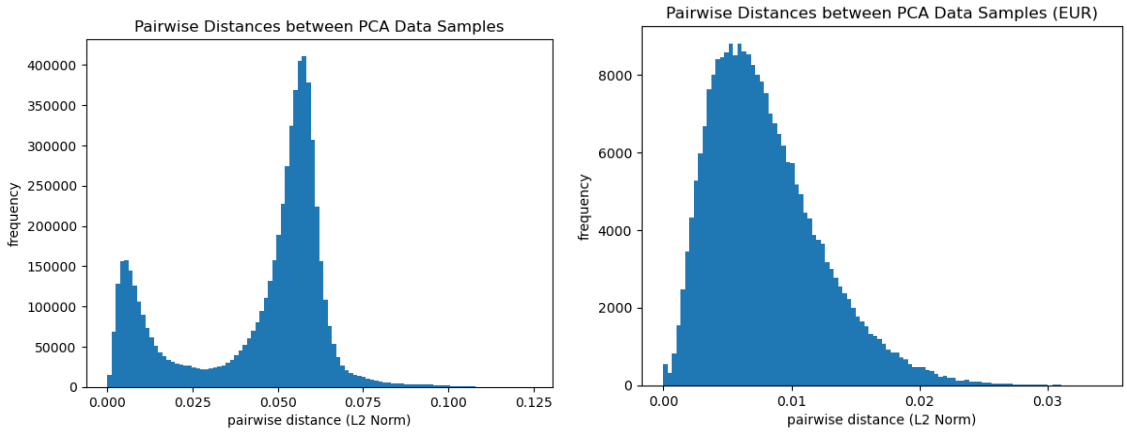
Figure 3: Truncated dendrogram (6 levels) from hierarchically clustering the LD-pruned SNP data. The subtrees (inferred clusters) are a large distance from their parent splits, indicating decent separation.

We deliberately tuned the clustering to select the optimal k in between 20 and 30 (true number of subpopulations was 27), and we ended up with a very low silhouette score and many subpopulation-level misassignments of points in the same population (Figure 2). The resulting silhouette score for optimal $k = 20$ was ~ 0.34, which agrees with the notion of many subpopulation clusters overlapping.

The best hierarchical methods (Ward linkage, average linkage) attained silhouette scores in the same range as k-means, settling on a tuned distance threshold yielding 5 clusters. The truncated dendrogram of the Ward clustering depicts this in figure 3. Having settled on such a low k, tuning method did not result in extra insight about subpopulation structure.

These results hint at some important characteristics of this data unaccounted for by these clustering methods. First and foremost, the distribution of pairwise distances between points is very different across the whole dataset compared to that within a single population. Ideally, the average intracluster distance that is distinct from and smaller than the average intercluster distance. This is the case across the whole dataset; but it is not the case within populations, as shown in figure 2.1.

Here we see that the distribution of pairwise distance between the data points in the entire dataset is bimodal; there is one peak (left) corresponding to intracluster distance, and another corresponding to intercluster distance. Within the `EUR` population the distribution of pairwise distance is



(a) pairwise distance distribution (whole dataset) (b) pairwise distance distribution (EUR)

Figure 4: Distributions of pairwise distances across $4a$) the whole dataset and $4b$) within the EUR population. The bimodal distribution of distances across the whole dataset makes the populations more separable than subpopulations within a single population, which have unimodally distributed pairwise distances.

unimodal, indicating indiscernible separation between subpopulations. This discrepancy is likely the reason both approaches failed to discern subpopulation structure, and further research on remedial preprocessing or clustering methods is needed.

2.2 Limitations & Future Research

This research had limitations which call for future research. Firstly, we used data from only chromosome 19. Data from more chromosomes could make subpopulation-level data more separable by adding more subpopulation-specific SNPs. Secondly, we used biologically/genetically uninformed euclidean distance on the PCA-transformed data. One could instead employ multi-dimensional scaling using custom, genetically informed pairwise distances. Thirdly, we used hierarchical clustering with a single distance threshold; the observed pairwise distances could warrant combining clusterings with multiple distance thresholds somehow, instead of taking a one-size-fits-all approach. Finally, one may want to use cluster quality score better suited for additional subcluster structure than silhouette score.

We also did not performance benchmark these methods for two reasons. First, hierarchical clustering methods have linkage-dependent asymptotic time complexities at least as slow as k-means. Secondly, we tuned different hyperparameters of k-means and hierarchical approaches. For these reasons, comparing these methods in this way would be useless at best and irresponsible at worst. We stuck with fairer comparison points like clustering quality and population structure captured.

References

[1] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[2] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[3] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[5] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[6] Robert R Sokal and Charles D Michener. A statistical method for evaluating systematic relationships. 1958.

[7] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.