1 Further details on clustering

This document provides additional details on the process of performing hierarchical agglomerative clustering, which is implemented in part using the `fastcluster` package.

1.1 Methodology

Generally, the process of agglomerative hierarchical clustering consists of merging groups of records together based on a distance criterion until all points are contained within one group, forming an inverse tree structure. From here it is possible to examine the tree and find the groupings that satisfy a certain linkage criterion (e.g., all records within each group returned are at most ϵ apart).

Formally, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^k$ define the set of records. Additionally, define a metric function $f(\mathbf{x}_i, \mathbf{x}_j)$ to evaluate the distance between two records and a linkage function $g(\mathcal{C}_p, \mathcal{C}_q)$, $\mathcal{C}_p \mathcal{C}_q \subset \mathcal{X}$ to determine the distance between two groups. Note that f is used as a kernel in g such that g is of the form $g(\mathcal{C}_p, \mathcal{C}_q) = h(f(\mathbf{x}_i, \mathbf{x}_j)) \, \forall \mathbf{x}_i \in \mathcal{C}_p$, $\mathbf{x}_j \in \mathcal{C}_q$. In a given iteration, for group $\mathcal{C}_i$, merge $\mathcal{C}_i$ with $\mathcal{C}^* = \underset{\mathcal{C}}{\operatorname{argmin}} \, g(\mathcal{C}_i, \mathcal{C})$. Continue this procedure for all groups until there is only one group remaining.

In our current implementation, we set f to be a weighted ℓ_2 norm of the form

$$f(\mathbf{x}_i, \mathbf{x}_j) = \Big((w(\mathbf{d}_{i,j}) \times \mathbf{d}_{i,j})^T (w(\mathbf{d}_{i,j}) \times \mathbf{d}_{i,j}) \Big)^{\frac{1}{2}} \quad \mathbf{d}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$$

where $w : \mathbb{R}^k \to \mathbb{R}^k$ is a weight function and $w(\mathbf{d}) \times \mathbf{d}$ denotes the element-wise product of the two vectors. We specify w as a function because of the flexibility it provides. We may wish to set the weight applied to the difference in the value of a tolerance variable between two records because we believe it may be a function of that difference. Otherwise, the function can just be a constant.

We set $g(\mathcal{C}_i, \mathcal{C}_j) = \max(f(\mathbf{x}_i, \mathbf{x}_j)) \, \forall \, \mathbf{x}_i \in \mathcal{C}_i$, $\mathbf{x}_j \in \mathcal{C}_j$. This corresponds to setting `method = 'complete'` in the `fastcluster` package. This linkage method is preferred because our intent is to establish an upper bound on the distance between any two records assigned to the same group.