

Appendix to
A generalized single linkage method for estimating the cluster
tree of a density

Werner Stuetzle *
Department of Statistics
University of Washington

Rebecca Nugent *
Department of Statistics
Carnegie Mellon University

* Supported by NSF grant DMS-9803226 and NSA grant 62-1942

Proof of Proposition 2

Proposition 2: The graph cluster tree of G is isomorphic to the single linkage dendrogram.

Lemma 1: Let (k, l) be an edge of G with weight $\hat{p}_{kl}^{(1)} > \lambda$. Then there is a path connecting \mathbf{x}_k and \mathbf{x}_l with maximum edge length $< 2/\lambda$.

Proof of Lemma 1: As $\hat{p}_{kl}^{(1)}$ is the minimum of the nearest neighbor density estimate over the line segment $[\mathbf{x}_k, \mathbf{x}_l]$, the assumption that $\hat{p}_{kl}^{(1)} > \lambda$ implies that the entire line segment $[\mathbf{x}_k, \mathbf{x}_l]$ is covered by spheres around the observations with radius $r = 1/\lambda$:

$$[\mathbf{x}_k, \mathbf{x}_l] \subset \bigcup_i S(\mathbf{x}_i, r).$$

Let

$$L_q = [\mathbf{a}_q, \mathbf{b}_q] = [\mathbf{x}_k, \mathbf{x}_l] \cup S(\mathbf{x}_q, r)$$

be the (possibly empty) intersection of the line segment $[\mathbf{x}_k, \mathbf{x}_l]$ with the sphere of radius r around \mathbf{x}_q . Without loss of generality assume that $d(\mathbf{x}_k, \mathbf{a}_q) \leq d(\mathbf{x}_k, \mathbf{b}_q)$. Choose $q_1 = k$. Because the L_q collectively cover $[\mathbf{x}_k, \mathbf{x}_l]$ there has to be a q_2 with $\mathbf{b}_{q_1} \in S(\mathbf{x}_{q_2}, r)$. Therefore, $d(\mathbf{x}_{q_1}, \mathbf{x}_{q_2}) < 2r$. Repeating this argument shows that there is a path connecting \mathbf{x}_k and \mathbf{x}_l with maximum edge length $< 2r = 2/\lambda$.

Lemma 2: The graph cluster tree of G and the cluster tree of the nearest neighbor density estimate are isomorphic.

Proof of Lemma 2: Let $\mathcal{X}_1, \dots, \mathcal{X}_k$ be vertex sets of the connected components of $G(\lambda)$. We will show that the connected components of $L(\lambda; \hat{p}^{(1)})$ are the sets

$$L_i = \bigcup_{\mathbf{x}_j \in \mathcal{X}_i} S(\mathbf{x}_j, 1/\lambda).$$

Suppose that L_i is connected. Then for any two vertices \mathbf{x}_j and \mathbf{x}_l in \mathcal{X}_i there exists a polyline connecting them with maximum edge length $< 2/\lambda$ and therefore minimum density $\hat{p}^{(1)} > \lambda$. This implies that \mathbf{x}_j and \mathbf{x}_l are in the same connected component of $G(\lambda)$.

On the other hand, suppose that \mathbf{x}_j and \mathbf{x}_l are in the same connected component of $G(\lambda)$. This implies that they are connected by a path with minimum edge weight $> \lambda$ and therefore maximum edge length $< 2/\lambda$ (Lemma 1), and hence are in the same connected component of $L(\lambda; \hat{p}^{(1)})$.

Proof of Proposition 2: According to Lemma 2, the graph cluster tree and the cluster tree of the nearest neighbor density estimate are isomorphic. On the other hand, Stuetzle (2003, Section 2) has shown that the cluster tree of the nearest neighbor density estimate is isomorphic to the single linkage dendrogram.