Appendices

A Wasserstein dimension selection

In order to compute the principal component scores $\zeta_1, \ldots, \zeta_n$, one must choose the dimension r. Traditionally and somewhat heuristically, this is done by finding the "elbow" in the scree plot of eigenvalues. The bias-variance tradeoff associated with choosing r was explored by [49], who suggested a data-splitting method of dimension selection for high-dimensional data using Wasserstein distances. Whilst this method may be more costly than the traditional "elbow" approach, it was empirically demonstrated in [49] to have superior performance.

Algorithm 2 Wasserstein PCA dimension selection [49]

Input: data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^p$.

1: **for** $r \in \{1, ..., \min(n, p)\}$ **do**
2: Let $\mathbf{V} \in \mathbb{R}^{p \times r}$ denote the matrix whose columns are orthonormal eigenvectors associated with the r largest eigenvalues of $\sum_{i=1}^{\lceil n/2 \rceil} \mathbf{Y}_i \mathbf{Y}_i^\top$
3: Orthogonally project $\mathbf{Y}_1, \ldots, \mathbf{Y}_{\lceil n/2 \rceil}$ onto the column space of $\mathbf{V}$, $\hat{\mathbf{Y}}_i := \mathbf{V}\mathbf{V}^\top \mathbf{Y}_i$
4: Compute Wasserstein distance d_r between $\hat{\mathbf{Y}}_i, \ldots, \hat{\mathbf{Y}}_{\lceil n/2 \rceil}$ and $\mathbf{Y}_{\lceil n/2 \rceil + 1}, \ldots, \mathbf{Y}_n$ (as point sets in $\mathbb{R}^p$)
5: **end for**

Output: selected dimension $\hat{r} = \text{argmin}\, \{d_r\}$.

We note that in practice, the eigenvectors appearing in this procedure could be computed sequentially as r grows, and to limit computational cost one might consider r only up to some $r_{\max} < \min(n, p)$.

B Implementation using scikit-learn

Algorithm 1 can be implemented using the Python module scikit-learn [38] via their `AgglomerativeClustering` class, using the standard 'average' linkage criterion and a custom metric function to compute the desired affinities. However, `AgglomerativeClustering` merges clusters based on minimum distance metric, whereas algorithm 1 merges according to maximum dot product. Therefore, the custom metric function we used in our implementation calculates all the pairwise dot products and subtracts them from the maximum. This transformation needs to then be rectified if accessing the merge heights.

All code released as part of this paper is under the MIT License and can be found at `https://github.com/anniegray52/dot_product_hierarchical`

C Proofs and supporting theoretical results

Lemma 2. *The height function $h(v) := \alpha(v, v)$ satisfies $h(v) \geq h(\text{Pa}_v)$ for all $v \in \mathcal{V}$ except the root.*

Proof.

$$\begin{aligned} h(v) &= \frac{1}{p}\mathbb{E}[\|\mathbf{X}(v) - \mathbf{X}(\text{Pa}_v) + \mathbf{X}(\text{Pa}_v)\|^2] \\ &= \frac{1}{p}\mathbb{E}[\|\mathbf{X}(v) - \mathbf{X}(\text{Pa}_v)\|^2] + 2\frac{1}{p}\mathbb{E}[\langle \mathbf{X}(v) - \mathbf{X}(\text{Pa}_v), \mathbf{X}(\text{Pa}_v)\rangle] + h(\text{Pa}_v) \geq h(\text{Pa}_v), \end{aligned}$$

where **A2** combined with the tower property and linearity of conditional expectation implies $\mathbb{E}[\langle \mathbf{X}(v) - \mathbf{X}(\text{Pa}_v), \mathbf{X}(\text{Pa}_v)\rangle] = 0$. □

Lemma 3. *For all $u, v \in \mathcal{V}$, $\alpha(u, v) \geq 0$.*

Proof. If $u = v$, $\alpha(u, v) \geq 0$ holds immediately from the definition of α in (2). For $u \neq v$ with most recent common ancestor w,

$$\begin{aligned}\mathbb{E}[\langle \mathbf{X}(u), \mathbf{X}(v)\rangle] &= \sum_{j=1}^{p} \mathbb{E}\left[\mathbb{E}[X_j(u)X_j(v)|X_j(w)]\right] \\ &= \sum_{j=1}^{p} \mathbb{E}\left[\mathbb{E}[X_j(u)|X_j(w)]\mathbb{E}[X_j(v)|X_j(w)]\right] \\ &= \sum_{j=1}^{p} \mathbb{E}[|X_j(w)|^2] \geq 0,\end{aligned}$$

where the second equality uses **A1** together with standard conditional independence aguments, and the third equality uses **A2**.

□

Proof of lemma 1. Let w be the most recent common ancestor of u and v. For each $j = 1, \ldots, p$, the property **A1** together with standard conditional independence arguments imply that $X_j(u)$ and $X_j(v)$ are conditionally independent given $X_j(w)$, and the property **A2** implies that $\mathbb{E}[X_j(u)|X_j(w)] = \mathbb{E}[X_j(v)|X_j(w)] = X_j(w)$. Therefore, by the tower property of conditional expectation,

$$\begin{aligned}\mathbb{E}[X_j(u)X_j(v)] &= \mathbb{E}\left[\mathbb{E}[X_j(u)X_j(v)|X_j(w)]\right] \\ &= \mathbb{E}\left[\mathbb{E}[X_j(u)|X_j(w)]\mathbb{E}[X_j(v)|X_j(w)]\right] \\ &= \mathbb{E}[X_j(w)^2].\end{aligned}$$

Hence, using the definitions of the merge height m, the height h and the affinity α,

$$m(u, v) = h(w) = \alpha(w, w) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}[X_j(w)^2] = \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}[X_j(u)X_j(v)] = \alpha(u, v),$$

which proves the first equality in the statement. The second equality is the definition of α.

For the third equality in the statement, we have

$$\begin{aligned}d(u, v) &= h(u) + h(v) - 2h(w) \\ &= \alpha(u, u) + \alpha(v, v) - 2\alpha(u, v) \\ &= \frac{1}{p}\mathbb{E}\left[\langle \mathbf{X}(u), \mathbf{X}(u)\rangle\right] + \frac{1}{p}\mathbb{E}\left[\langle \mathbf{X}(v), \mathbf{X}(v)\rangle\right] - 2\frac{1}{p}\mathbb{E}\left[\langle \mathbf{X}(u), \mathbf{X}(v)\rangle\right] \\ &= \frac{1}{p}\mathbb{E}\left[\|\mathbf{X}(u) - \mathbf{X}(v)\|^2\right],\end{aligned}$$

where the first equality uses the definition of d, and the second equality uses the definition of h and $h(w) = m(u, v) = \alpha(u, v)$. □

C.1 Proof of Theorem 1

The following lemma establishes an identity concerning the affinities computed in algorithm 1 which will be used in the proof of theorem 1.

Lemma 4. *Let P_m, $m \geq 0$, be the sequence of partitions of $[n]$ constructed in algorithm 1. Then for any $m \geq 0$,*

$$\hat{\alpha}(u, v) = \frac{1}{|u||v|}\sum_{i \in u, j \in v} \hat{\alpha}(i, j), \quad \textit{for all distinct pairs } u, v \in P_m. \tag{9}$$

Proof. The proof is by induction on m. With $m = 0$, (9) holds immediately since $P_0 = \{\{1\}, \ldots, \{n\}\}$. Now suppose (9) holds at step m. Then for any distinct pair $w, w' \in P_{m+1}$,

either w or w' is the result of merging two elements of P_m, or w and w' are both elements of P_m. In the latter case the induction hypothesis immediately implies:

$$\hat{\alpha}(w, w') = \frac{1}{|w||w'|} \sum_{i \in w, j \in w'} \hat{\alpha}(i, j).$$

In the case that w or w' is the result of a merge, suppose w.l.o.g. that $w = u \cup v$ for some $u, v \in P_m$ and $w' \in P_m$. Then by definition of $\hat{\alpha}$ in algorithm 1,

$$\begin{aligned}
\hat{\alpha}(w, w') &= \frac{|u|}{|w|}\hat{\alpha}(u, w') + \frac{|v|}{|w|}\hat{\alpha}(v, w') \\
&= \frac{|u|}{|w|}\frac{1}{|u||w'|} \sum_{i \in u, j \in w'} \hat{\alpha}(i, j) + \frac{|v|}{|w|}\frac{1}{|v||w'|} \sum_{i \in v, j \in w'} \hat{\alpha}(i, j) \\
&= \frac{1}{|w||w'|} \sum_{i \in w, j \in w'} \hat{\alpha}(i, j),
\end{aligned}$$

where the final equality uses $w = u \cup v$. The induction hypothesis thus holds at step $m + 1$. □

The following proposition establishes the validity of the height function constructed in algorithm 1. Some of the arguments used in this proof are qualitatively similar to those used to study reducible linkage functions by, e.g., Sumengen et al. [46], see also historical references therein.

Proposition 1. *With $\hat{\mathcal{V}}$ the vertex set and $\hat{h}$ the height function constructed in algorithm 1 with any symmetric, real-valued input $\hat{\alpha}(\cdot, \cdot)$, it holds that $\hat{h}(v) \geq \hat{h}(\mathrm{Pa}_v)$ for all vertices $v \in \hat{\mathcal{V}}$ except the root.*

Proof. The required inequality $\hat{h}(v) \geq \hat{h}(\mathrm{Pa}_v)$ holds immediately for all the leaf vertices $v \in P_0 = \{\{1\}, \ldots, \{n\}\}$ by the definition of $\hat{h}$ in algorithm 1. All the remaining vertices in the output tree, i.e., those in $\hat{\mathcal{V}} \setminus P_0$, are formed by merges over the course of the algorithm. For $m \geq 0$ let $w_m = u_m \cup v_m$ denote the vertex formed by merging some $u_m, v_m \in P_m$. Then $w_m = \mathrm{Pa}_{u_m}$ and $w_m = \mathrm{Pa}_{v_m}$. Each u_m is either a member of P_0 or equal to $w_{m'}$ for some $m' < m$. The same is true of each v_m. It therefore suffices to show that $\hat{h}(w_m) \geq \hat{h}(w_{m+1})$ for $m \geq 0$, where by definition in the algorithm, $\hat{h}(w_m) = \hat{\alpha}(u_m, v_m)$. Also by definition in the algorithm, $\hat{h}(w_{m+1})$ is the largest pairwise affinity between elements of P_{m+1}. Our objective therefore is to upper-bound this largest affinity and compare it to $\hat{h}(w_m) = \hat{\alpha}(u_m, v_m)$.

The affinity between $w_m = u_m \cup v_m$ and any other element w' of P_{m+1} (which must also be an element of P_m) is, by definition in the algorithm,

$$\begin{aligned}
\hat{\alpha}(w_m, w') &= \frac{|u_m|}{|u_m| + |v_m|}\hat{\alpha}(u_m, w') + \frac{|v_m|}{|u_m| + |v_m|}\hat{\alpha}(v_m, w') \\
&\leq \max\{\hat{\alpha}(u_m, w'), \hat{\alpha}(v_m, w')\} \\
&\leq \hat{\alpha}(u_m, v_m),
\end{aligned}$$

where the last inequality holds because u_m, v_m, by definition, have the largest affinity amongst all elements of P_m. For the same reason, the affinity between any two distinct elements of P_{m+1} neither of which is w_m (and therefore both of which are elements of P_m) is upper-bounded by $\hat{\alpha}(u_m, v_m)$. We have therefore established $\hat{h}(w_m) = \hat{\alpha}(u_m, v_m) \geq \hat{h}(w_{m+1})$ as required, and this completes the proof.

□

Proof of theorem 1. Let us introduce some definitions used throughout the proof.

$$M \coloneqq \max_{i \neq j} |\hat{\alpha}(i, j) - \alpha(z_i, z_j)|. \tag{10}$$

We arbitrarily chose and then fix $i, j \in [n]$ with $i \neq j$, and define

$$H \coloneqq m(z_i, z_j), \qquad \hat{H} \coloneqq \hat{m}(i, j). \tag{11}$$

Let u denote the most recent common ancestor of the leaf vertices $\{i\}$ and $\{j\}$ in $\hat{\mathcal{D}}$ and let $m \geq 1$ denote the step of the algorithm at which u is created by a merge, that is $m = \min\{m' \geq 1 : u \in P_{m'}\}$. We note that by construction, u is equal to the union of all leaf vertices with ancestor u, and by definition of $\hat{h}$ in algorithm 1 $\hat{h}(u) = \hat{H}$.

Let v denote the most recent common ancestor of z_i and z_j in $\mathcal{D}$, which has height $h(v) = H$.

Lower bound on $m(z_i, z_j) - \hat{m}(i, j)$. There is no partition of u into two non-empty sets $A, B \subseteq [n]$ such that $\hat{\alpha}(k, l) < \hat{H}$ for all $k \in A$ and $l \in B$. We prove this by contradiction. Suppose that such a partition exists. There must be a step $m' \leq m$ at which some $A' \subseteq A$ is merged some $B' \subseteq B$. The vertex w formed by this merge would have height

$$\begin{aligned}\hat{h}(w) &= \hat{\alpha}(A', B') \\ &= \frac{1}{|A'||B'|} \sum_{k \in A', l \in B'} \hat{\alpha}(k, l) < \hat{H} = \hat{h}(u),\end{aligned}$$

where the first equality is the definition of $\hat{h}(w)$ in the algorithm and the second equality holds by lemma 4. However, in this construction u is an ancestor of w, and $\hat{h}(w) < \hat{h}(u)$ therefore contradicts the result of proposition 1.

As a device to be used in the next step of the proof, consider an undirected graph with vertex set u, in which there is an edge between two vertices k and l if and only if $\hat{\alpha}(k, l) \geq \hat{H}$. Then, because there is no partition as established above, this graph must be connected. Now consider a second undirected graph, also with vertex set u, in which there is any edge between two vertices k and l if and only if $\alpha(z_k, z_l) \geq \hat{H} - M$. Due to the definition of M in (10), any edge in the first graph is an edge in the second, so the second graph is connected too. Let k, l, and ℓ be any distinct members of u. Using the fact established in lemma 1 that $\alpha(z_k, z_l)$ and $\alpha(z_l, z_\ell)$ are respectively the merge heights in $\mathcal{D}$ between z_k and z_l, and z_l and z_ℓ, it can be seen that if there are edges between k and l and between l and ℓ in the second graph, there must also be an edge in that graph between k and ℓ. Combined with the connectedness, this implies that the second graph is complete, so that $\alpha(z_k, z_l) \geq \hat{H} - M$ for all distinct $k, l \in u$. In particular $\alpha(z_i, z_j) \geq \hat{H} - M$, and since $m(z_i, z_j) = \alpha(z_i, z_j)$, we find

$$m(z_i, z_j) - \hat{m}(i, j) \geq -M. \tag{12}$$

Upper bound on $m(z_i, z_j) - \hat{m}(i, j)$. Let $S_v = \{i \in [n] : z_i = v \text{ or } z_i \text{ has ancestor } v \text{ in } \mathcal{D}\}$. For $k, l \in S_v$, lemma 1 tells us $\alpha(z_k, z_l)$ is the merge height between z_k and z_l, so $\alpha(z_k, z_l) \geq H$. Using (10), we therefore have

$$\hat{\alpha}(k, l) \geq H - M, \quad \forall k, l \in S_v. \tag{13}$$

It follows from the definition of $\hat{h}$ in the algorithm that if $S_v = [n]$, the heights of all vertices in $\hat{\mathcal{D}}$ are greater than or equal to $H - M$. This implies $\hat{H} \geq H - M$. In summary, we have shown that when $S_v = [n]$,

$$m(z_i, z_j) - \hat{m}(i, j) \leq M. \tag{14}$$

It remains to consider the case $S_v \neq [n]$. The proof of the same upper bound (14) in this case is more involved. In summary, we need to establish that the most recent common ancestor of $\{i\}$ and $\{j\}$ in $\hat{\mathcal{D}}$ has height at least $H - M$. The main idea of the proof is to consider the latest step of the algorithm at which a vertex with height at least $H - M$ is formed by a merge, and show the partition formed by this merge contains the most recent common ancestor of $\{i\}$ and $\{j\}$, or an ancestor thereof.

To this end let m^* denote the latest step in algorithm 1 at which the vertex formed, w^*, has height greater than or equal to $H - M$. To see that m^* must exist, notice

$$\max_{k \neq l \in [n]} \hat{\alpha}(k, l) \geq \alpha(z_i, z_j) - M, \tag{15}$$

by definition of M in (10). Combined with the definition of $\hat{h}$ in algorithm 1, the vertex formed by the merge at step 1 of the algorithm therefore has height greater than or equal to $H - M$. Therefore m^* is indeed well-defined.

Our next objective is to show that the partition P_{m^*} formed at step $m^\star$ contains an element which itself contains both i and j. We proceed by establishing some facts about S_v and P_{m^*}.

Let $\bar{S}_v := [n] \setminus S_v$. For $k \in S_v$, $l \in \bar{S}_v$, v cannot be an ancestor of z_l, by lemma 1 $\alpha(z_k, z_l)$ is the merge height of z_k and z_l, and b is the minimum branch length in $\mathcal{D}$, so we have $\alpha(z_k, z_l) \leq H - b$. From (10) we then find

$$\hat{\alpha}(k, l) \leq H - b + M, \quad \forall\, k \in S_v, l \in \bar{S}_v. \tag{16}$$

We claim that no element of P_{m^*} can contain both an element of S_v and an element of $\bar{S}_v$. We prove this claim by contradiction. If such an element of P_{m^*} did exist, there would be a step $m' \leq m^*$ at which some $A' \subseteq S_v$ is merged with some $B' \subseteq \bar{S}_v$. But the vertex w' formed by this merge would be assigned height $\hat{h}(w') = \hat{\alpha}(A', B') \leq H - b + M < H - M$, where the first inequality uses lemma 4 and (16), and the second inequality uses the assumption of the theorem that $M < b/2$. Recalling the definition of $w^\star$ we have $\hat{h}(w^*) \geq H - M$. We therefore see that w^* is an ancestor of w' with $\hat{h}(w^*) > \hat{h}(w')$, contradicting the result of proposition 1.

Consider the elements of P_{m^*}, denoted A and B, which contain i and j respectively. We claim that $A = B$. We prove this claim by contradiction. Suppose $A \neq B$. As established in the previous paragraph, neither A nor B can contain an element of $\bar{S}_v$. Therefore, using lemma 4 and (13),

$$\hat{\alpha}(A, B) = \frac{1}{|A||B|} \sum_{k \in A, l \in B} \hat{\alpha}(k, l) \geq H - M.$$

Again using the established fact that no element of P_{m^*} can contain both an element of S_v and an element of $\bar{S}_v$, m^* cannot be the final step of the algorithm, since that would require $P_{m^\star} = \{[n]\}$. Therefore $\hat{\alpha}(A, B)$ is one of the affinities which algorithm 1 would maximise over at step $m^* + 1$, so the height of the vertex formed by a merge at step $m^* + 1$ would be greater than or equal to $H - M$, which contradicts the definition of m^*. Thus we have proved there exists an element of P_{m^*} which contains both i and j. This element must be the most recent common ancestor of $\{i\}$ and $\{j\}$, or an ancestor thereof. Also, this element must have been formed by a merge at a step less than or equal to $m^\star$ and so must have height greater than or equal to $H - M$. Invoking proposition 1 we have thus established $\hat{H} \geq H - M$. In summary, in the case $S_v \neq [n]$, we have shown

$$m(z_i, z_j) - \hat{m}(i, j) \leq M. \tag{17}$$

Combining the lower bound (12) with the upper bounds (14), (17) and the fact that i, j were chosen arbitrarily, completes the proof.

□

C.2 Supporting material and proof for Theorem 2

Definitions and interpretation for assumptions A3 and A5

We recall the definition of φ-mixing from, e.g., [39]. For a sequence of random variables $\{\xi_j; j \geq 1\}$, define:

$$\varphi(k) := \sup_{j \geq 1} \sup_{A \in \mathcal{F}_1^j, B \in \mathcal{F}_{j+k}^\infty, \mathbb{P}(A) > 0} |\mathbb{P}(B|A) - \mathbb{P}(B)|\,.$$

where $\mathcal{F}_i^j$ is the σ-algebra generated by $\xi_i, \ldots, \xi_j$. Then $\{\xi_j; j \geq 1\}$ is said to be φ-mixing if $\varphi(k) \searrow 0$ as $k \to \infty$.

To interpret assumption A5 notice

$$\mathbb{E}[\|\mathbf{S}(Z_i)\mathbf{E}_i\|^2 | Z_1, \ldots, Z_n] \leq \|\mathbf{S}(Z_i)\|_{\text{op}}^2 \mathbb{E}[\|\mathbf{E}_i\|^2] \leq \max_{v \in \mathcal{Z}} \|\mathbf{S}(v)\|_{\text{op}}^2\, p\, \mathbb{E}[|\mathbf{E}_{11}|^2],$$

where the first inequality uses the independence of $\mathbf{E}_i$ and Z_i and the second inequality uses the fact that the elements of the vectors $\mathbf{E}_i$ are i.i.d. Since $\mathbf{Y}_i - \mathbf{X}(Z_i) = \mathbf{S}(Z_i)\mathbf{E}_i$, A5 thus implies $\mathbb{E}[\|\mathbf{Y}_i - \mathbf{X}(Z_i)\|^2] \in O(p)$ as $p \to \infty$, which can be viewed as a natural growth rate since p is the dimension of the disturbance vector $\mathbf{Y}_i - \mathbf{X}(Z_i)$. In the proof of proposition 2 below, A5 is used in a similar manner to control dot products of the form $\langle \mathbf{Y}_i - \mathbf{X}_i, \mathbf{X}_j \rangle$ and $\langle \mathbf{Y}_i - \mathbf{X}_i, \mathbf{Y}_j - \mathbf{X}_j \rangle$.

Proof of Theorem 2. For the first claim of the theorem, proposition 2 combined with the tower property of conditional expectation imply that for any $\delta > 0$,

$$\mathbb{P}\left(\max_{1\le i<j\le n} \left|p^{-1}\left\langle \mathbf{Y}_i, \mathbf{Y}_j\right\rangle - \alpha(Z_i, Z_j)\right| \ge \delta\right) \le \frac{1}{\delta^q}\frac{1}{p^{q/2}}\frac{n(n-1)}{2} C(q,\varphi) M(q, \mathbf{X}, \mathbf{E}, \mathbf{S}), \quad (18)$$

from which (6) follows.

The second claim of the theorem is in essence a corollary to [49][Thm 1]. A little work is needed to map the setting of the present work on to the setting of Whiteley et al. [49][Thm 1]. To see the connection, we endow the finite set $\mathcal{Z}$ in the present work with the discrete metric: $d_{\mathcal{Z}}(u,v) := 0$ for $u \neq v$, and $d_{\mathcal{Z}}(v,v) = 0$. Then $(\mathcal{Z}, d_{\mathcal{Z}})$ is a compact metric space, and in the setting specified in the statement of theorem 2 where **A3** is strengthened to independence, $s = p$ and $\mathbf{S}(v) = \sigma \mathbf{I}_p$ for all $v \in \mathcal{Z}$, the variables $\mathbf{Y}_1 \ldots, \mathbf{Y}_n$; $\{\mathbf{X}(v), v \in \mathcal{Z}\}$; $\mathbf{E}_1, \ldots, \mathbf{E}_n$ exactly follow the Latent Metric Model of Whiteley et al. [49].

Moreover, according to the description in section 2.1, the variables $Z_1, \ldots, Z_n$ are i.i.d. according to a probability distribution supported on $\mathcal{Z}$. As in [49], by Mercer's theorem there exists a feature map $\phi : \mathcal{Z} \to \mathbb{R}^r$ associated with this probability distribution, such that $\langle \phi(u), \phi(v)\rangle = \alpha(u,v)$, for $u, v \in \mathcal{Z}$. Here r, as in **A6**, is the rank of the matrix with elements $\alpha(u,v)$, which is at most $\mathcal{Z}$.

Theorem 1 of [49] in this context implies there exists a random orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r\times r}$ such that

$$\max_{i\in[n]} \left\| p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\| \in O_{\mathbb{P}}\left(\sqrt{\frac{nr}{p}} + \sqrt{\frac{r}{n}}\right). \quad (19)$$

Consider the bound:

$$\begin{aligned}
|\hat{\alpha}_{\text{pca}}(i,j) - \alpha(Z_i, Z_j)| &= \left|\frac{1}{p}\langle \zeta_i, \zeta_j\rangle - \langle \phi(Z_i), \phi(Z_j)\rangle\right| \\
&\le \left|\left\langle p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i), p^{-1/2}\mathbf{Q}\zeta_j\right\rangle\right| \\
&\quad + \left|\left\langle \phi(Z_i), p^{-1/2}\mathbf{Q}\zeta_j - \phi(Z_j)\right\rangle\right| \\
&\le \left\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\| \left(\left\|p^{-1/2}\mathbf{Q}\zeta_j - \phi(Z_j)\right\| + \|\phi(Z_j)\|\right) \\
&\quad + \|\phi(Z_i)\| \left\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\|,
\end{aligned}$$

where orthogonality of $\mathbf{Q}$ has been used, and the final inequality uses Cauchy-Schwarz and the triangle inequality for the $\|\cdot\|$ norm. Combining the above estimate with (19), the bound:

$$\begin{aligned}
\max_{i\in[n]} \|\phi(Z_i)\|^2 \le \max_{v\in\mathcal{Z}} \|\phi(v)\|^2 = \max_{v\in\mathcal{Z}} \alpha(v,v) \\
\le \sup_{j\ge 1}\max_{v\in\mathcal{Z}} \mathbb{E}[|X_j(v)|^2] \le \sup_{j\ge 1}\max_{v\in\mathcal{Z}} \mathbb{E}[|X_j(v)|^{2q}]^{1/q} \quad (20)
\end{aligned}$$

and **A4** completes the proof of the second claim of the theorem.

□

Proposition 2. *Assume the model in section 2.1 satisfies assumptions A3-A5, and let φ and q be as in A3 and A4. Then there exists a constant $C(q,\varphi)$ depending only on q and φ such that for any $\delta > 0$,*

$$\mathbb{P}\left(\max_{1\le i<j\le n} \left|p^{-1}\left\langle \mathbf{Y}_i, \mathbf{Y}_j\right\rangle - \alpha(Z_i, Z_j)\right| \ge \delta \,\middle|\, Z_1, \ldots, Z_n\right) \le \frac{1}{\delta^q}\frac{1}{p^{q/2}}\frac{n(n-1)}{2} C(q,\varphi) M(q, \mathbf{X}, \mathbf{E}, \mathbf{S}) \quad (21)$$

where

$$M(q, \mathbf{X}, \mathbf{E}, \mathbf{S}) \coloneqq \sup_{k\geq 1} \max_{v\in\mathcal{Z}} \mathbb{E}\left[|X_k(v)|^{2q}\right] + \mathbb{E}\left[|\mathbf{E}_{11}|^q\right] \left(\sup_{p\geq 1} \max_{v\in\mathcal{Z}} \|\mathbf{S}(v)\|_{\text{op}}^q\right) \sup_{k\geq 1} \max_{v\in\mathcal{Z}} \mathbb{E}\left[|X_k(v)|^q|\right] + \mathbb{E}\left[|\mathbf{E}_{11}|^{2q}\right] \sup_{p\geq 1} \max_{v\in\mathcal{Z}} \|\mathbf{S}(v)\|_{\text{op}}^{2q}.$$

Proof. Fix any i, j such that $1 \leq i < j \leq n$. Consider the decomposition:

$$p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j\rangle - \alpha(Z_i, Z_j) = \sum_{k=1}^{4} \Delta_k$$

where

$$\begin{aligned}
\Delta_1 &\coloneqq p^{-1}\langle \mathbf{X}(Z_i), \mathbf{X}(Z_j)\rangle - \alpha(Z_i, Z_j)\\
\Delta_2 &\coloneqq p^{-1}\langle \mathbf{X}(Z_i), \mathbf{S}(Z_j)\mathbf{E}_j\rangle\\
\Delta_3 &\coloneqq p^{-1}\langle \mathbf{X}(Z_j), \mathbf{S}(Z_i)\mathbf{E}_i\rangle\\
\Delta_4 &\coloneqq p^{-1}\langle \mathbf{S}(Z_i)\mathbf{E}_i, \mathbf{S}(Z_j)\mathbf{E}_j\rangle
\end{aligned}$$

The proof proceeds by bounding $\mathbb{E}[|\Delta_k|^q|Z_1, \ldots, Z_n]$ for $k = 1, \ldots, 4$. Writing Δ_1 as

$$\Delta_1 = \frac{1}{p}\sum_{k=1}^{p} \Delta_{1,k}, \qquad \Delta_{1,k} \coloneqq X_k(Z_i)X_k(Z_j) - \mathbb{E}\left[X_k(Z_i)X_k(Z_j)|\, Z_1, \ldots, Z_n\right].$$

we see that Δ_1 is a sum p random variables each of which is conditionally mean zero given $Z_1, \ldots, Z_n$. Noting that the two collections of random variables $\{Z_1, \ldots, Z_n\}$ and $\{\mathbf{X}(v); v \in \mathcal{V}\}$ are independent (as per the description of the model in section 2.1), under assumption A3 we may apply a moment inequality for φ-mixing random variables [51][Lemma 1.7] to show that there exists a constant $C_1(q, \varphi)$ depending only on q, φ such that

$$\begin{aligned}
&\mathbb{E}\left[|\Delta_1|^q|\, Z_1, \ldots, Z_n\right]\\
&\leq C_1(q,\varphi)\left\{\frac{1}{p^q}\sum_{k=1}^{p}\mathbb{E}\left[|\Delta_{1,k}|^q|\, Z_1, \ldots, Z_n\right] + \left(\frac{1}{p^2}\sum_{k=1}^{p}\mathbb{E}\left[|\Delta_{1,k}|^2\right|\, Z_1, \ldots, Z_n\right]\right)^{q/2}\right\}\\
&\leq C_1(q,\varphi)\left\{\frac{1}{p^q}\sum_{k=1}^{p}\mathbb{E}\left[|\Delta_{1,k}|^q|\, Z_1, \ldots, Z_n\right] + \frac{1}{p^{q/2}}\frac{1}{p}\sum_{k=1}^{p}\mathbb{E}\left[|\Delta_{1,k}|^q|\, Z_1, \ldots, Z_n\right]\right\}\\
&\leq 2C_1(q,\varphi)\frac{1}{p^{q/2}}\sup_{k\geq 1}\mathbb{E}\left[|\Delta_{1,k}|^q|\, Z_1, \ldots, Z_n\right]\\
&\leq 2^{q+1}C_1(q,\varphi)\frac{1}{p^{q/2}}\sup_{k\geq 1}\max_{v\in\mathcal{Z}}\mathbb{E}\left[|X_k(v)|^{2q}\right],
\end{aligned} \tag{22}$$

where second inequality holds by two applications of Jensen's inequality and $q \geq 2$, and the final inequality uses the fact that for $a, b \geq 0$, $(a+b)^q \leq 2^{q-1}(a^q + b^q)$, the Cauchy-Schwartz inequality, and the independence of $\{Z_1, \ldots, Z_n\}$ and $\{\mathbf{X}(v); v \in \mathcal{V}\}$.

For Δ_2, we have

$$\Delta_2 \coloneqq \frac{1}{p}\sum_{k=1}^{p}\Delta_{2,k}, \qquad \Delta_{2,k} \coloneqq [\mathbf{S}(Z_j)^\top \mathbf{X}(Z_i)]_k \mathbf{E}_{jk},$$

where $[\cdot]_k$ denotes the kth element of a vector. Since the three collections of random variables, $\{Z_1, \ldots, Z_n\}$, $\{\mathbf{X}(v); v \in \mathcal{Z}\}$ and $\{\mathbf{E}_1, \ldots, \mathbf{E}_n\}$ are mutually independent, and the elements of each vector $\mathbf{E}_j \in \mathbb{R}^p$ are mean zero and independent, we see that given $\{Z_1, \ldots, Z_n\}$ and $\{\mathbf{X}(v); v \in \mathcal{V}\}$, Δ_2 is a simple average of conditionally independent and conditionally mean-zero

random variables. Applying the Marcinkiewicz–Zygmund inequality we find there exists a constant $C_2(q)$ depending only on q such that

$$\mathbb{E}\left[\,|\Delta_2|^q \,\middle|\, Z_1,\dots,Z_n,\mathbf{X}(v);v\in\mathcal{Z}\right] \le C_2(q)\mathbb{E}\left[\left|\frac{1}{p^2}\sum_{k=1}^{p}|\Delta_{2,k}|^2\right|^{q/2}\,\middle|\, Z_1,\dots,Z_n,\mathbf{X}(v);v\in\mathcal{Z}\right]. \quad (23)$$

Noting that $q\ge 2$ and applying Minkowski's inequality to the r.h.s. of (23), then using the independence of $\{Z_1,\dots,Z_n\}$, $\{\mathbf{X}(v);v\in\mathcal{Z}\}$ and $\{\mathbf{E}_1,\dots,\mathbf{E}_n\}$ and the i.i.d. nature of the elements of the vector $\mathbf{E}_j$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{p^2}\sum_{k=1}^{p}|\Delta_{2,k}|^2\right|^{q/2}\,\middle|\, Z_1,\dots,Z_n,\mathbf{X}(v);v\in\mathcal{Z}\right]^{2/q}\\
&\le \frac{1}{p^2}\sum_{k=1}^{p}\mathbb{E}\left[\,|\Delta_{2,k}|^q\,|\,Z_1,\dots,Z_n,\mathbf{X}(v);v\in\mathcal{Z}\right]^{2/q}\\
&= \frac{1}{p^2}\sum_{k=1}^{p}\mathbb{E}\left[\left|[\mathbf{S}(Z_j)^\top\mathbf{X}(Z_i)]_k\right|^q|\mathbf{E}_{jk}|^q\,\middle|\, Z_1,\dots,Z_n,\mathbf{X}(v);v\in\mathcal{Z}\right]^{2/q}\\
&= \frac{1}{p^2}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]^{2/q}\sum_{k=1}^{p}\left|[\mathbf{S}(Z_j)^\top\mathbf{X}(Z_i)]_k\right|^2\\
&= \frac{1}{p^2}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]^{2/q}\left\|\mathbf{S}(Z_j)^\top\mathbf{X}(Z_i)\right\|^2\\
&\le \frac{1}{p^2}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]^{2/q}\max_{v\in\mathcal{Z}}\|\mathbf{S}(v)\|_{\mathrm{op}}^2\,\|\mathbf{X}(Z_i)\|^2.
\end{aligned}$$

Substituting into (23) and using the tower property of conditional expectation we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\,|\Delta_2|^q\,|\,Z_1,\dots,Z_n\right]\\
&\le \frac{1}{p^{q/2}}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]\max_{v\in\mathcal{Z}}\|\mathbf{S}(v)\|_{\mathrm{op}}^q\,\frac{1}{p^{q/2}}\mathbb{E}\left[\,\|\mathbf{X}(Z_i)\|^q\,|\,Z_1,\dots,Z_n\right]\\
&= \frac{1}{p^{q/2}}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]\max_{v\in\mathcal{Z}}\|\mathbf{S}(v)\|_{\mathrm{op}}^q\,\mathbb{E}\left[\left(\frac{1}{p}\sum_{k=1}^{p}|X_k(Z_j)|^2\right)^{q/2}\,\middle|\, Z_1,\dots,Z_n\right]\\
&\le \frac{1}{p^{q/2}}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]\max_{v\in\mathcal{Z}}\|\mathbf{S}(v)\|_{\mathrm{op}}^q\,\mathbb{E}\left[\frac{1}{p}\sum_{k=1}^{p}|X_k(Z_j)|^q\,\middle|\, Z_1,\dots,Z_n\right]\\
&\le \frac{1}{p^{q/2}}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]\max_{v\in\mathcal{Z}}\|\mathbf{S}(v)\|_{\mathrm{op}}^q\sup_{k\ge1}\max_{v\in\mathcal{Z}}\mathbb{E}\left[\,|X_k(v)|^q|\right] \quad (24)
\end{aligned}$$

where the second inequality holds by Jensen's inequality (recall $q\ge 2$). Since the r.h.s. of (24) does not depend on i or j, the same bound holds with Δ_2 on the l.h.s. replaced by Δ_3.

Turning to Δ_4, we have

$$\Delta_4 := \frac{1}{p}\left\langle\mathbf{S}(Z_i)\mathbf{E}_i,\mathbf{S}(Z_j)\mathbf{E}_j\right\rangle = \frac{1}{p}\sum_{1\le k,\ell\le p}\Delta_{4,k,\ell},\qquad \Delta_{4,k,\ell} := \mathbf{E}_{ik}\mathbf{E}_{j\ell}[\mathbf{S}(Z_i)^\top\mathbf{S}(Z_j)]_{k\ell}.$$

Noting that $i\ne j$, and that the elements of $\mathbf{E}_i$ and $\mathbf{E}_j$ are independent, identically distributed, and mean zero, we see that Δ_4 is a sum of p^2 random variables which are all conditionally mean zero and conditionally independent given $Z_1,\dots,Z_n$. The Marcinkiewicz–Zygmund inequality gives:

$$\mathbb{E}\left[\,|\Delta_4|^q\,|\,Z_1,\dots,Z_n\right] \le C_2(q)\mathbb{E}\left[\left|\frac{1}{p^2}\sum_{1\le k,\ell\le s}^{p}|\Delta_{4,k,\ell}|^2\right|^{q/2}\,\middle|\, Z_1,\dots,Z_n\right]. \quad (25)$$

Applying Minkowski's inequality to the r.h.s. of (25),

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{p^2}\sum_{1\le k,\ell\le p}|\Delta_{4,k,\ell}|^2\right|^{q/2}\middle| Z_1,\dots,Z_n\right]^{2/q}\\
&\le \frac{1}{p^2}\sum_{1\le k,\ell\le p}\mathbb{E}\left[|\Delta_{4,k,\ell}|^q\middle| Z_1,\dots,Z_n\right]^{2/q}\\
&= \frac{1}{p^2}\sum_{1\le k,\ell\le p}\mathbb{E}\left[|\mathbf{E}_{ik}|^q\,|\mathbf{E}_{j\ell}|^q\left|[\mathbf{S}(Z_i)^\top\mathbf{S}(Z_j)]_{k\ell}\right|^q\middle| Z_1,\dots,Z_n\right]^{2/q}\\
&= \frac{1}{p^2}\mathbb{E}\left[|\mathbf{E}_{11}|^q\right]^{4/q}\sum_{1\le k,\ell\le p}\left|[\mathbf{S}(Z_i)^\top\mathbf{S}(Z_j)]_{k\ell}\right|^2\\
&\le \frac{1}{p^2}\mathbb{E}\left[|\mathbf{E}_{11}|^{2q}\right]^{2/q}\max_{u,v\in\mathcal{Z}}\|\mathbf{S}(u)^\top\mathbf{S}(v)\|_{\mathrm{F}}^2,
\end{aligned}$$

where the final inequality holds by Jensen's inequality. Substituting back into (25) and using $\|\mathbf{S}(u)^\top\mathbf{S}(v)\|_{\mathrm{F}} \le p^{1/2}\|\mathbf{S}(u)^\top\mathbf{S}(v)\|_{\mathrm{op}} \le p^{1/2}\|\mathbf{S}(u)\|_{\mathrm{op}}\|\mathbf{S}(v)\|_{\mathrm{op}}$, we obtain:

$$\mathbb{E}\left[|\Delta_4|^q\middle| Z_1,\dots,Z_n\right] \le C_2(q)\frac{1}{p^{q/2}}\mathbb{E}\left[|\mathbf{E}_{11}|^{2q}\right]\max_{u\in\mathcal{Z}}\|\mathbf{S}(u)\|_{\mathrm{op}}^{2q}. \tag{26}$$

Combining (22), (24) and (26) using the fact that for $a, b \ge 0$, $(a+b)^q \le 2^{q-1}(a^q+b^q)$, we find that there exists a constant $C(q,\varphi)$ depending only on q and φ such that

$$\begin{aligned}
&\mathbb{E}\left[\left|p^{-1}\langle\mathbf{Y}_i,\mathbf{Y}_j\rangle - \alpha(Z_i,Z_j)\right|^q\middle| Z_1,\dots,Z_n\right]\\
&\le C(q,\varphi)\frac{1}{p^{q/2}}M(q,\mathbf{X},\mathbf{E},\mathbf{S}),
\end{aligned}$$

where $M(q,\mathbf{X},\mathbf{E},\mathbf{S})$ is defined in the statement of the proposition and is finite by assumptions **A4** and **A5**. By Markov's inequality, for any $\delta \ge 0$,

$$\mathbb{P}\left(\left|p^{-1}\langle\mathbf{Y}_i,\mathbf{Y}_j\rangle - \alpha(Z_i,Z_j)\right| \ge \delta\middle| Z_1,\dots,Z_n\right) \le \frac{1}{\delta^q}C(q,\varphi)\frac{1}{p^{q/2}}M(q,\mathbf{X},\mathbf{E},\mathbf{S}) \tag{27}$$

and the proof is completed by a union bound:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le i<j\le n}\left|p^{-1}\langle\mathbf{Y}_i,\mathbf{Y}_j\rangle - \alpha(Z_i,Z_j)\right| < \delta\middle| Z_1,\dots,Z_n\right)\\
&= \mathbb{P}\left(\bigcap_{1\le i<j\le n}\left|p^{-1}\langle\mathbf{Y}_i,\mathbf{Y}_j\rangle - \alpha(Z_i,Z_j)\right| < \delta\middle| Z_1,\dots,Z_n\right)\\
&= 1-\mathbb{P}\left(\bigcup_{1\le i<j\le n}\left|p^{-1}\langle\mathbf{Y}_i,\mathbf{Y}_j\rangle - \alpha(Z_i,Z_j)\right| \ge \delta\middle| Z_1,\dots,Z_n\right)\\
&\ge 1-\frac{n(n-1)}{2}\frac{1}{\delta^q}C(q,\varphi)\frac{1}{p^{q/2}}M(q,\mathbf{X},\mathbf{E},\mathbf{S}).
\end{aligned}$$

□

C.3 Interpretation of merge heights and exact tree recovery

Here we expand on the discussion in section 3.3 and provide further interpretation of merge heights and algorithm 1. In particular our aim is to clarify in what circumstances algorithm 1 will asymptotically correctly recover underlying tree structure. For ease of exposition throughout section C.3 we assume that $\mathcal{Z}$ are the leaf vertices of $\mathcal{T}$.

As a preliminary we note the following corollary to theorem 1: assuming $b > 0$, if one takes as input to algorithm 1 the true merge heights, i.e. (up to bijective relabelling of leaf vertices)

$\hat{\alpha}(\cdot,\cdot) \coloneqq m(\cdot,\cdot) = \alpha(\cdot,\cdot)$, where $n = |\mathcal{Z}|$, then theorem 1 implies that algorithm 1 outputs a dendrogram $\mathcal{D}$ whose merge heights $\hat{m}(\cdot,\cdot)$ are equal to $m(\cdot,\cdot)$ (up to bijective relabeling over vertices). This clarifies that with knowledge of $m(\cdot,\cdot)$, algorithm 1) constructs a dendrogram which has $m(\cdot,\cdot)$ as its merge heights.

We now ask for more: if once again $m(\cdot,\cdot)$ is taken as input to algorithm 1, under what conditions is the output tree $\hat{\mathcal{T}}$ equal to $\mathcal{T}$ (upto bijective relabelling of vertices)? We claim this holds when $\mathcal{T}$ is a binary tree and that all its non-leaf nodes have different heights. We provide a sketch proof of this claim, since a complete proof involves many tedious and notationally cumbersome details.

To remove the need for repeated considerations of relabelling, suppose $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ is given, then w.l.o.g. relabel the leaf vertices of $\mathcal{T}$ as $\{1\}, \ldots, \{|\mathcal{Z}|\}$ and relabel each non-leaf vertex to be the union of its children. Thus each vertex is some subset of $[|\mathcal{Z}|]$.

Now assume that $\mathcal{T}$ is a binary tree and that all its non-leaf nodes have different heights. Note that $|\mathcal{V}| = 2|\mathcal{Z}| - 1$, i.e., there are $|\mathcal{Z}| - 1$ non-leaf vertices. The tree $\mathcal{T}$ is uniquely characterized by a sequence of partitions $\tilde{P}_0, \ldots, \tilde{P}_{|\mathcal{Z}|-1}$ where $\tilde{P}_0 \coloneqq \{\{1\}, \ldots, \{|\mathcal{Z}|\}\}$, and for $m = 1, \ldots, |\mathcal{Z}| - 1$, $\tilde{P}_m$ is constructed from $\tilde{P}_{m-1}$ by merging the two elements of $\tilde{P}_{m-1}$ whose most recent common ancestor is the mth highest non-leaf vertex (which is uniquely defined since we are assuming no two non-leaf vertices have equal heights).

To see that in this situation algorithm 1, with $\hat{\alpha}(\cdot,\cdot) \coloneqq m(\cdot,\cdot)$ and $n = |\mathcal{Z}|$ as input, performs exact recovery of the tree, i.e., $\hat{\mathcal{T}} = \mathcal{T}$, it suffices to notice that the sequence of partitions $P_0, \ldots, P_{|\mathcal{Z}|-1}$ constructed by algorithm 1 uniquely characterizes $\hat{\mathcal{T}}$, and moreover $(\tilde{P}_0, \ldots, \tilde{P}_{|\mathcal{Z}|-1}) = (P_0, \ldots, P_{|\mathcal{Z}|-1})$. The details of this last equality involve simple but tedious substitutions of $m(\cdot,\cdot)$ in place of $\hat{\alpha}(\cdot,\cdot)$ in algorithm 1, so are omitted.

D Further details of numerical experiments and data preparation

All real datasets used are publicly available under the CC0: Public domain license. Further, all experiments were run locally on a laptop with an integrated GPU (Intel UHD Graphics 620).

D.1 Simulated data

For each $v \in \mathcal{V}$, $X_1(v), \ldots, X_p(v)$ are independent and identically distributed Gaussian random variables with:

$$
\begin{aligned}
X_j(1) &\sim N(X_j(6), 5),\\
X_j(2) &\sim N(X_j(6), 2),\\
X_j(3) &\sim N(X_j(6), 2),\\
X_j(4) &\sim N(X_j(7), 0.5),\\
X_j(5) &\sim N(X_j(7), 7),\\
X_j(6) &\sim N(X_j(8), 2),\\
X_j(7) &\sim N(X_j(8), 1),\\
X_j(8) &\sim N(0, 1),
\end{aligned}
$$

for $j = 1, \ldots, p$.

D.2 20 Newsgroups

The dataset originates from [29], however, the version used is the one available in the Python package 'scikit-learn' [38]. Each document is pre-processed in the following way: generic stopwords and e-mail addresses are removed, and words are lemmatised. The processed documents are then converted into a matrix of TF-IDF features. Labels can be found on the 20 Newsgroups website `http://qwone.com/~jason/20Newsgroups/`, but are mainly intuitive from the title of labels, with full stops separating levels of hierarchy. When using PCA a dimension of $r = 34$ was selected by the method described in appendix A.

The following numerical results complement those in the main part of the paper.

Table 2: Kendall τ_b ranking performance measure, for Algorithm 1 and the 20 Newsgroups data set. The mean Kendall τ_b correlation coefficient is reported alongside the standard error (numerical value shown is the standard error$\times 10^3$). This numerical results are plotted in figure 4 below.

Data	Input	$p = 500$	$p = 1000$	$p = 5000$	$p = 7500$	$p = 12818$
Newsgroups	$\mathbf{Y}_{1:n}$	0.026 (0.55)	0.016 (1.0)	0.13 (2.2)	0.17 (2.5)	0.26 (2.9)
	$\zeta_{1:n}$	0.017 (0.72)	0.047 (1.2)	0.12 (1.9)	0.15 (2.5)	0.24 (2.6)

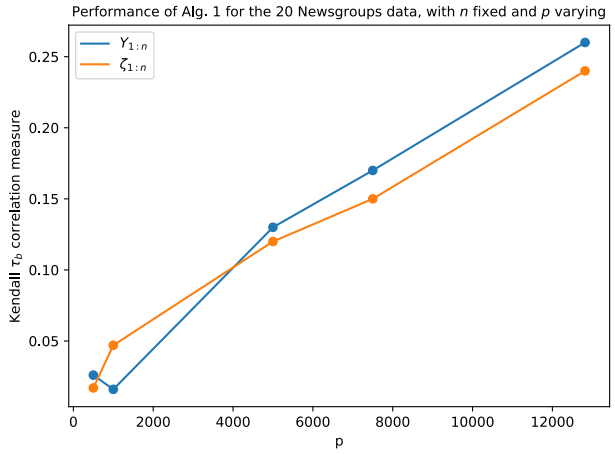

Figure 4: Performance of Algorithm 1 for the 20 Newsgroups data set as a function of number of TF-IDF features, p, with n fixed. See table 1 for numerical values and standard errors.

D.3 Zebrafish gene counts

These data were collected over a 24-hour period (timestamps available in the data), however, the temporal aspect of the data was ignored when selecting a sub-sample. To process the data, we followed the steps in [47] which are the standard steps in the popular SCANPY [50] – a Python package for analysing single-cell gene-expression data to process the data. This involves filtering out genes that are present in less than 3 cells or are highly variable, taking the logarithm, scaling and regressing out the effects of the total counts per cell. When using PCA a dimension of $r = 29$ was selected by the method described in appendix A.

D.4 Amazon reviews

The pre-processing of this dataset is similar to that of the Newsgroup data except e-mail addresses are no longer removed. Some data points had a label of 'unknown' in the third level of the hierarchy, these were removed from the dataset. In addition, reviews that are two words or less are not included. When using PCA a dimension of $r = 22$ was selected by the method described in appendix A.

D.5 S&P 500 stock data

This dataset was found through the paper [8] with the authors code used to process the data. The labels follow the Global Industry Classification Standard and can be found here: [3]. The return on day i of each stock is calculated by

$$\text{return}_i = \frac{p_i^{cl} - p_i^{op}}{p_i^{op}},$$

where p_i^{op} and p_i^{cl} is the respective opening and closing price on day i. When using PCA a dimension of $r = 10$ was used as selected by the method described in appendix A.

D.6 Additional method comparison

Table 3 reports additional method comparison results, complementing those in table 1 which concerned only the average linkage function (noting UPGMA is equivalent to using the average linkage function with Euclidean distances). In table 3 we also compare to Euclidean and cosine distances paired with complete and single linkage functions. To aid comparison, the first column (average linkage with the dot product) is the same as in table 1. In general, using complete or single linkage performs worse for both Euclidean and cosine distances. The only notable exception being a slight improvement on the simulated dataset.

Table 3: Kendall τ_b ranking performance measure. For the dot product method, i.e., algorithm 1, $\mathbf{Y}_{1:n}$ as input corresponds to using $\hat{\alpha}_{\text{data}}$, and $\zeta_{1:n}$ corresponds to $\hat{\alpha}_{\text{pca}}$. The mean Kendall τ_b correlation coefficient is reported alongside the standard error (numerical value shown is the standard error $\times 10^3$).

Data	Input	Average linkage	Complete linkage		Single linkage	
		Dot product	Euclidean	Cosine	Euclidean	Cosine
Newsgroups	$\mathbf{Y}_{1:n}$	0.26 (2.9)	0.022 (0.87)	-0.010 (0.88)	-0.0025 (0.62)	-0.0025 (0.62)
	$\zeta_{1:n}$	0.24 (2.6)	0.0041 (1.2)	0.036 (1.2)	-0.016 (2.0)	0.067 (1.5)
Zebrafish	$\mathbf{Y}_{1:n}$	0.34 (3.4)	0.15 (2.2)	0.24 (3.2)	0.023 (3.0)	0.032 (2.9)
	$\zeta_{1:n}$	0.34 (3.4)	0.17 (2.0)	0.30 (3.4)	0.12 (2.8)	0.15 (3.2)
Reviews	$\mathbf{Y}_{1:n}$	0.15 (2.5)	0.019 (0.90)	0.023 (1.0)	0.0013 (0.81)	0.0013 (0.81)
	$\zeta_{1:n}$	0.14 (2.4)	0.058 (1.5)	0.063 (1.8)	0.015 (1.2)	0.038 (1.0)
S&P 500	$\mathbf{Y}_{1:n}$	0.34 (10)	0.33 (10)	0.33 (10)	0.17 (10)	0.17 (10)
	$\zeta_{1:n}$	0.36 (9.4)	0.32 (10)	0.31 (10)	0.36 (13)	0.39 (12)
Simulated	$\mathbf{Y}_{1:n}$	0.86 (1)	0.55 (8.7)	0.84 (2.0)	0.55 (8.7)	0.84 (2.0)
	$\zeta_{1:n}$	0.86 (1)	0.55 (8.7)	0.84 (2.0)	0.55 (8.7)	0.84 (2.0)

Table 4: Kendall τ_b ranking performance measure. The mean Kendall τ_b correlation coefficient is reported alongside the standard error (numerical value shown is the standard error$\times 10^3$).

Data	Input	UPGMA with dot product *dissimilarity*	UPGMA with Manhattan distance
Newsgroups	$\mathbf{Y}_{1:n}$	-0.0053 (0.24)	-0.0099 (1.3)
	$\zeta_{1:n}$	0.0029 (0.33)	0.052 (1.6)
Zebrafish	$\mathbf{Y}_{1:n}$	0.0012 (0.13)	0.16 (2.4)
	$\zeta_{1:n}$	0.00046 (0.12)	0.050 (2.8)
Reviews	$\mathbf{Y}_{1:n}$	-0.0005 (0.29)	0.0018 (0.44)
	$\zeta_{1:n}$	-0.0015 (0.41)	0.061 (1.3)
S&P 500	$\mathbf{Y}_{1:n}$	0.0026 (7.7)	0.37 (9.4)
	$\zeta_{1:n}$	0.0028 (7.5)	0.39 (11)
Simulated	$\mathbf{Y}_{1:n}$	-0.0026 (1.6)	0.55 (8.7)
	$\zeta_{1:n}$	-0.0023 (1.8)	0.84 (2)

E Understanding agglomerative clustering with Euclidean or cosine distances in our framework

E.1 Quantifying dissimilarity using Euclidean distance

The first step of many standard variants of agglomerative clustering such as UPGMA and Ward's method is to find and merge the pair of data vectors which are closest to each other in Euclidean distance. From the elementary identity:

$$\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 = \|\mathbf{Y}_i\|^2 + \|\mathbf{Y}_j\|^2 - 2\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle,$$

we see that, in general, this is not equivalent to finding the pair with largest dot product, because of the presence of the terms $\|\mathbf{Y}_i\|^2$ and $\|\mathbf{Y}_j\|^2$. For some further insight in to how this relates to our

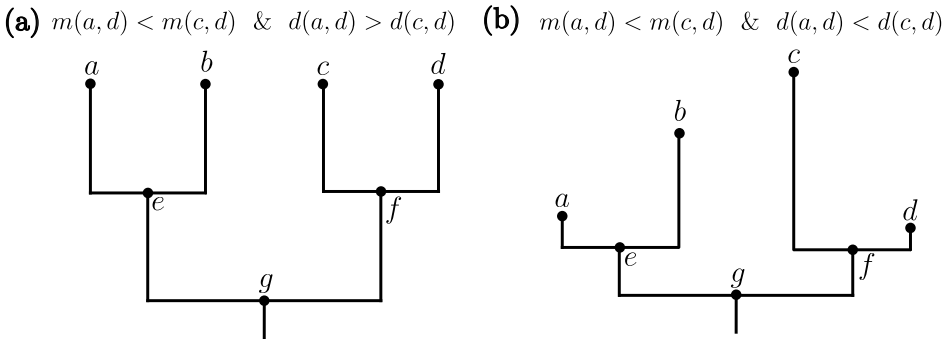

Figure 5: Illustration of how maximising merge height $m(\cdot,\cdot)$ may or may not be equivalent to minimising distance $d(\cdot,\cdot)$, depending on the geometry of the dendrogram. (a) Equivalence holds (b) Equivalence does not hold.

modelling and theoretical analysis framework, it is revealing to the consider the idealised case of choosing to merge by maximising merge height $m(\cdot,\cdot)$ versus minimising $d(\cdot,\cdot)$ (recall the identities for m and d established in lemma 1). Figure 5 shows two simple scenarios in which geometry of the dendrogram has an impact on whether or not maximising merge height $m(\cdot,\cdot)$ is equivalent to minimising $d(\cdot,\cdot)$. From this example we see that, in situations where some branch lengths are disproportionately large, minimising $d(\cdot,\cdot)$ will have different results to maximising $m(\cdot,\cdot)$.

UPGMA and Ward's method differ in their linkage functions, and so differ in the clusters they create in practice after their respective first algorithmic steps. UPGMA uses average linkage to combine Euclidean distances, and there does not seem to be a mathematically simple connection between this and algorithm 1, except to say that in general it will return different results. Ward's method merges the pair of clusters which results in the minimum increase in within-cluster variance. When clusters contain equal numbers of samples, this increase is equal to the squared Euclidean distance between the clusters' respective centroids.

E.2 Agglomerative clustering with cosine distance is equivalent to an instance of algorithm 1

The cosine 'distance' between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ is:

$$d_{\cos}(i,j) \coloneqq 1 - \frac{\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle}{\|\mathbf{Y}_i\| \|\mathbf{Y}_j\|}.$$

In table 1 we show results for standard agglomerative clustering using $d_{\cos}$ as a measure of dissimilarity, combined with average linkage. At each iteration, this algorithm works by merging the pair of data-vectors/clusters which are closest with respect to $d_{\cos}(\cdot,\cdot)$, say u and v merged to form w, with dissimilarities between w and the existing data-vectors/clusters computed according to the average linkage function:

$$d_{\cos}(w,\cdot) \coloneqq \frac{|u|}{|w|} d_{\cos}(u,\cdot) + \frac{|v|}{|w|} d_{\cos}(v,\cdot). \tag{28}$$

This procedure can be seen to be equivalent to algorithm 1 with input $\hat{\alpha}(\cdot,\cdot) = 1 - d_{\cos}(\cdot,\cdot)$. Indeed maximising $1 - d_{\cos}(\cdot,\cdot)$ is clearly equivalent to minimizing $d_{\cos}(\cdot,\cdot)$, and with $\hat{\alpha}(\cdot,\cdot) = 1 - d_{\cos}(\cdot,\cdot)$ the affinity computation at line 6 of algorithm 1 is:

$$\begin{aligned}
\hat{\alpha}(w,\cdot) &\coloneqq \frac{|u|}{|w|}\hat{\alpha}(u,\cdot) + \frac{|v|}{|w|}\hat{\alpha}(v,\cdot) \\
&= \frac{|u|}{|w|}\left[1 - d_{\cos}(u,\cdot)\right] + \frac{|v|}{|w|}\left[1 - d_{\cos}(v,\cdot)\right] \\
&= \frac{|u|+|v|}{|w|} - \frac{|u|}{|w|} d_{\cos}(u,\cdot) - \frac{|u|}{|w|} d_{\cos}(v,\cdot) \\
&= 1 - \left[\frac{|u|}{|w|} d_{\cos}(u,\cdot) + \frac{|u|}{|w|} d_{\cos}(v,\cdot)\right],
\end{aligned}$$

where on the r.h.s. of the final equality we recognise (28).

E.3 Using cosine similarity as an affinity measure removes multiplicative noise

Building from the algorithmic equivalence identified in section E.2, we now address the theoretical performance of agglomerative clustering with cosine distance in our modelling framework.

Intuitively, cosine distance is used in situations where the magnitudes of data vectors are thought not to convey useful information about dissimilarity. To formalise this idea, we consider a variation of our statistical model from section 2.1 in which:

- $\mathcal{Z}$ are the leaf vertices of $\mathcal{T}$, $|\mathcal{Z}| = n$, and we take these vertices to be labelled $\mathcal{Z} = \{1, \dots, n\}$
- $\mathbf{X}$ is as in section 2.1, properties **A1** and **A2** hold, and it is assumed that for all $v \in \mathcal{Z}$, $p^{-1}\mathbb{E}[\|\mathbf{X}(v)\|^2] = 1$.
- the additive model (1) is replaced by a multiplicative noise model:
$$\mathbf{Y}_i = \gamma_i \mathbf{X}(i), \tag{29}$$
where $\gamma_i > 0$ are all strictly positive random scalars, independent of other variables, but otherwise arbitrary.

The interpretation of this model is that the expected square magnitude of the data vector $\mathbf{Y}_i$ is entirely determined by γ_i, indeed we have

$$\frac{1}{p}\mathbb{E}[\|\mathbf{Y}_i\|^2|\gamma_i] = \gamma_i^2 \frac{1}{p}\mathbb{E}[\|\mathbf{X}(i)\|^2] = \gamma_i^2 h(i) = \gamma_i^2,$$

where h is as in section 2.1. We note that in this multiplicative model, the random vectors $\mathbf{E}_i$ and matrices $\mathbf{S}(v)$ from section 2.1 play no role, and one can view the random variables $Z_1, \dots, Z_n$ as being replaced by constants $Z_i = i$, rather than being i.i.d.

Now define:

$$\hat{\alpha}_{\cos}(i,j) \coloneqq \frac{\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle}{\|\mathbf{Y}_i\|\|\mathbf{Y}_j\|} = 1 - d_{\cos}(i,j).$$

Theorem 3. *Assume that the model specified in section E.3 satisfies A3 and for some $q \geq 2$, $\sup_{j \geq 1} \max_{v \in \mathcal{Z}} \mathbb{E}[|X_j(v)|^{2q}] < \infty$. Then*

$$\max_{i,j \in [n], i \neq j} |\alpha(i,j) - \hat{\alpha}_{\cos}(i,j)| \in O_{\mathbb{P}}\left(\frac{n^{2/q}}{\sqrt{p}}\right).$$

Proof. The main ideas of the proof are that the multiplicative factors γ_i,γ_j in the numerator $\hat{\alpha}_{\cos}(i,j)$ cancel out with those in the denominator, and combined with the condition $p^{-1}\mathbb{E}[\|\mathbf{X}(v)\|^2] = 1$ we may then establish that $\alpha(i,j)$ and $\hat{\alpha}_{\cos}(i,j)$ are probabilistically close using similar arguments to those in the proof of proposition 2.

Consider the decomposition:

$$\begin{aligned}
\alpha(i,j) - \hat{\alpha}_{\cos}(i,j) &= \alpha(i,j) - \frac{p^{-1}\langle \mathbf{X}(i), \mathbf{X}(j) \rangle}{p^{-1/2}\|\mathbf{X}(i)\| p^{-1/2}\|\mathbf{X}(j)\|} \\
&= \alpha(i,j) - \frac{1}{p}\langle \mathbf{X}(i), \mathbf{X}(j) \rangle && (30) \\
&\quad + \frac{\frac{1}{p}\langle \mathbf{X}(i), \mathbf{X}(j) \rangle}{p^{-1/2}\|\mathbf{X}(i)\| p^{-1/2}\|\mathbf{X}(j)\|} \left[p^{-1/2}\|\mathbf{X}(i)\| p^{-1/2}\|\mathbf{X}(j)\| - 1 \right]. && (31)
\end{aligned}$$

Applying the Cauchy-Schwartz inequality, and adding and subtracting $p^{-1/2}\|\mathbf{X}(j)\|$ and 1 in the final term of this decomposition leads to:

$$\begin{aligned}
&\max_{i,j \in [n], i \neq j} |\alpha(i,j) - \hat{\alpha}_{\cos}(i,j)| \\
&\leq \max_{i,j \in [n], i \neq j} \left| \alpha(i,j) - \frac{1}{p}\langle \mathbf{X}(i), \mathbf{X}(j) \rangle \right| && (32) \\
&\quad + \max_{i \in [n]} \left| p^{-1/2}\|\mathbf{X}(i)\| - 1 \right| && (33) \\
&\quad + \left(1 + \max_{i \in [n]} \left| p^{-1/2}\|\mathbf{X}(i)\| - 1 \right| \right) \max_{j \in [n]} \left| p^{-1/2}\|\mathbf{X}(j)\| - 1 \right|. && (34)
\end{aligned}$$

The proof proceeds by arguing that the term (32) is in $O_{\mathbb{P}}\left(n^{2/q}/\sqrt{p}\right)$, and the terms (33) and (34) are in $O_{\mathbb{P}}\left(n^{1/q}/\sqrt{p}\right)$, where we note that this asymptotic concerns the limit as $n^{2/q}/\sqrt{p} \to 0$.

In order to analyse the term (32), let $\tilde{\mathbb{P}}$ denote the probability law of the additive model for $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ in equation (1) in section 2.1, in the case that $\mathbf{S}(v) = 0$ for all $v \in \mathcal{Z}$. Let $\tilde{\mathbb{E}}$ denote the associated expectation. Then for any $\delta > 0$ and $i, j \in [n]$, $i \neq j$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\alpha(i,j) - \frac{1}{p}\langle \mathbf{X}(i), \mathbf{X}(j)\rangle\right| > \delta\right) \\
&= \tilde{\mathbb{P}}\left(\left|\frac{1}{p}\langle \mathbf{Y}_i, \mathbf{Y}_j\rangle - \alpha(Z_i, Z_j)\right| > \delta \middle| Z_1 = 1, \ldots, Z_n = n\right) \\
&= \frac{\tilde{\mathbb{E}}\left[\mathbb{I}\{Z_1 = 1, \ldots, Z_n = n\}\tilde{\mathbb{P}}\left(\left|\frac{1}{p}\langle \mathbf{Y}_i, \mathbf{Y}_j\rangle - \alpha(Z_i, Z_j)\right| > \delta \middle| Z_1, \ldots, Z_n\right)\right]}{\tilde{\mathbb{P}}(\{Z_1 = 1, \ldots, Z_n = n\})}.
\end{aligned}$$

The conditional probability on the r.h.s of the final equality can be upper bounded using the inequality (27) in the proof of proposition 2, and combined with the same union bound argument used immediately after (27), this establishes (32) is in $O_{\mathbb{P}}\left(n^{2/q}/\sqrt{p}\right)$ as required.

To show that (33) and (34) are in $O_{\mathbb{P}}\left(n^{1/q}/\sqrt{p}\right)$, it suffices to show that $\max_{i\in[n]}\left|p^{-1/2}\|\mathbf{X}(i)\| - 1\right|$ is in $O_{\mathbb{P}}\left(n^{1/q}/\sqrt{p}\right)$. By re-arranging the equality: $(|a| - 1)(|a| + 1) = |a|^2 - 1$, we have

$$\left|p^{-1/2}\|\mathbf{X}(i)\| - 1\right| \leq \left|p^{-1}\|\mathbf{X}(i)\|^2 - 1\right| = \left|p^{-1}\sum_{j=1}^{p}\Delta_j(i)\right|, \tag{35}$$

where $\Delta_j(i) := |X_j(i)|^2 - 1$. Thus under the model from section E.3, $p^{-1}\sum_{j=1}^{p}\Delta_j(i)$ is an average of p mean-zero random variables, and by the same arguments as in the proof of proposition 2 under the mixing assumption **A3** and the assumption of the theorem that $\sup_{j\geq 1}\max_{v\in\mathcal{Z}}\mathbb{E}[|X_j(v)|^{2q}] < \infty$, combined with a union bound, we have

$$\max_{i\in[n]}\left|p^{-1}\|\mathbf{X}(i)\|^2 - 1\right| \in O_{\mathbb{P}}(n^{1/q}/\sqrt{p}).$$

Together with (35) this implies $\max_{i\in[n]}\left|p^{-1/2}\|\mathbf{X}(i)\| - 1\right|$ is in $O_{\mathbb{P}}\left(n^{1/q}/\sqrt{p}\right)$ as required, and that completes the proof.

□

E.4 Limitations of our modelling assumptions and failings of dot-product affinities

As noted in section 5, algorithm 1 is motivated and theoretically justified by our modelling assumptions, laid out in sections 2-3 and E.3. If these assumptions are not well matched to data in practice, then algorithm 1 may not perform well.

The proof of lemma 3 shows that, as a consequence of the conditional independence and martingale-like assumptions, **A1** and **A2**, $\alpha(u, v) \geq 0$ for all u, v. By theorems 2 and 3, $\hat{\alpha}_{\text{data}}$, $\hat{\alpha}_{\text{pca}}$ approximate α (at the vertices $\mathcal{Z}$) under the additive model from section 2.1, and $\hat{\alpha}_{\cos}$ approximates α under the multiplicative model from section E.3. Therefore if in practice $\hat{\alpha}_{\text{data}}(i,j)$, $\hat{\alpha}_{\text{pca}}(i,j)$ or $\hat{\alpha}_{\cos}(i,j)$ are found to take non-negligible negative values for some pairs i, j, that is an indication that our modelling assumptions may not be appropriate.

Even if the values taken by $\hat{\alpha}_{\text{data}}(i,j)$, $\hat{\alpha}_{\text{pca}}(i,j)$ or $\hat{\alpha}_{\cos}(i,j)$ are all positive in practice, there could be other failings of our modelling assumptions. As an academic but revealing example, if instead of (1) or (29), the data were to actually follow a combined additive *and* multiplicative noise model, i.e.,

$$\mathbf{Y}_i = \gamma_i\mathbf{X}(Z_i) + \mathbf{S}(Z_i)\mathbf{E}_i,$$

then in general neither $\hat{\alpha}_{\text{data}}$, $\hat{\alpha}_{\text{pca}}$ nor $\hat{\alpha}_{\cos}$ would approximate α as $p \to \infty$.