Principle Component Trees and their Persistent Homology: Technical Appendix

Anonymous submission

Abstract

Low dimensional models like PCA are often used to simplify complex datasets by learning a *single* approximating subspace. This paradigm has expanded to *union of subspaces* models, like those learned by subspace clustering. In this paper, we present *Principle Component Trees* (PCTs), a graph structure that generalizes these ideas to identify mixtures of components that together describe the subspace structure of high-dimensional datasets. Each node in a PCT corresponds to a principle component of the data, and the edges between nodes indicate the components that must be *mixed* to produce a subspace that approximates a portion of the data. In order to construct PCTs, we propose two angle-distribution hypothesis tests to detect subspace clusters in the data. To analyze, compare, and select the best PCT model, we define two persistent homology measures that describe their shape. We show our construction yields two key properties of PCTs, namely ancestral orthogonality and non-decreasing singular values. Our main theoretical results show that learning PCTs reduces to PCA under multivariate normality, and that PCTs are efficient parameterizations of intersecting union of subspaces. Finally, we use PCTs to analyze neural network latent space, word embeddings, and reference image datasets.

1 Appendix Outline.

In this appendix, we present material to supplement the AAAI submission, ”Principle Component Trees and Their Persistent Homology”. While we include the abstract for reference, this appendix is not made to stand alone. We present 3 sections of supplementary material.

- Practical Details / Considerations for the implementation of our PCT construction algorithm.
- More detailed proofs of our theoretical results listed in the paper.
- P-Values for experimental comparisons.

2 Practical considerations for PCT construction

In the main paper, we stated:

> For simplicity and interpretability, we leave three important implementation details about this and the following test to the technical appendix. 1) In order to reasonably compare angle distributions to those of hyper*spheres*, we first whiten $\mathbf{X}$ before taking the angle distribution. 2) This *explicit* whitening makes the expected angle distribution $\mathbb{P}_D(C)$ dependent on N, so we used monte-carlo simulation to estimate $\mathbb{P}_D(C)$ for $N < 100D$. 3) Whenever comparing the subspace clusteredness of two data matrices, we first project both to a common dimension.

We will now examine these details in order.

Data Pre-Whitening

In section 4, we compare an observed angle distribution with an expected distribution under uniformity on the hypersphere. This CDF for absolute cosine similarities was given in (Thordsen and Schubert 2022) as:

$$\mathbb{P}_D(C) = 2\ \text{Beta}_{\text{CDF}}\left(\tfrac{1+C}{2};\ \ \alpha = \tfrac{D-1}{2}, \beta = \tfrac{D-1}{2}\right) - 1.$$

Unfortunately, this expected distribution of angles will only properly apply *isotropic distributions*. The expexted distribution of angles will be different for anisotropic distributions. To isolate differences in the observed $\angle\mathbf{X}$ and theoretical $\mathbb{P}_D(C)$ that occur because of subspace clustering behavior and not an isotropic covariance, we whiten $\mathbf{X}$ such that it's diagonal matrix of singular values will be exactly $\mathbf{I}_D$.

Figure 1 demonstrates why this is necessary. All three distributions display roughly the same anisotropic covariance before whitening, but only two could be reasonably called subspace clustered (1b and 1c). Only after whitening the data can we make an “apples-to-apples” comparison of angle distribution.

Monte Carlo Estimation of $\mathbb{P}_D(C)$

Unfortunately, the *explicit* whitening we discuss in the previous section has a side effect. Because whitening makes the covariance **exactly** isotropic, it makes points ”over-perpendicular” to what would occur randomly when points are sampled uniformly from a hypershere. In the extreme case, when we whiten D points in $\mathbb{R}^D$, all points become exactly perpendicular to each other. See Figure **??** for an example of this in $\mathbb{R}^2$.

Because of this over-whitening, the expected distribution of pairwise angles is no longer dependent only on the

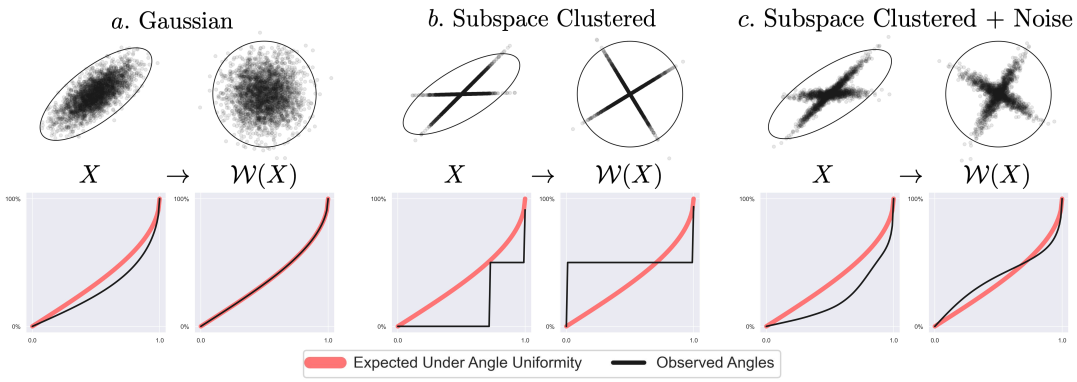


Figure 1: Effect of whitening on observed pairwise angle distribution.

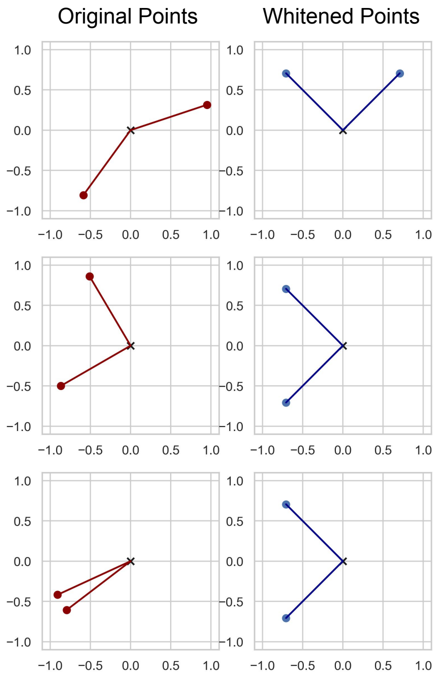


Figure 2: Data whitening makes any 2 points in $\mathbb{R}^2$ exactly perpendicular to each other.

data dimension D, but also the number of points N. We denote this new, datasize-dependent expected distribution $\mathbb{P}_{D,N}(C)$. This means we can't use the existing theoretical CDF of angle distribution when dealing with pre-whitened data; it is only the *asymptotic* distribution. Deriving the exact, non-asymptotic angle distribution would be a great direction for future work, but we were unable to find this expression exactly. Instead, we turned to Monte-Carlo simulation.

More specifically, we estimate $\mathbb{P}_{D,N}(C)$ for combinations of D and N from $D = 2$ to $D = 8$, and for $N = D$ to $N = 500D$, taking only 100 values of N in this range. For each combination, we ran enough simulations to collect $100,000,000$ whitened angles, and cached empirical CDF

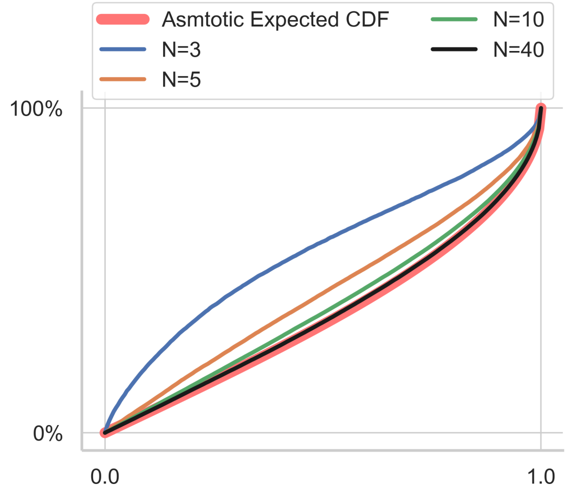


Figure 3: Expected angle CDF for different data sizes. Ambient space is $\mathbb{R}^2$

of those angles as $\mathbb{P}_{D,N}(C)$ for later use. Thankfully, the domain of $\mathbb{P}_{D,N}(C)$ is only $[0, 1]$, so we didn't need to estimate small tails of the distribution. Figure 3 shows 4 such Monte-Carlo CDFs and the asymptotic CDF.

Whenever we needed to perform a Cramer-Von-Mises test to compare $\mathbf{X}$ to $\mathbb{P}_{D,N}(\hat{C})$, we load the closest cached $\mathbb{P}_{D,N}(C)$ and compare our $\hat{\mathbb{P}}_{D,N}(C)$ to it.

Projection to a common dimension for comparison.

When discussing the test of subspace intersection, we state that our test $\texttt{INTERSECT-TEST}(\mathbf{X}, \mathbf{v})$ returns $\texttt{true}$ if

$$\texttt{CLUSTER-TEST}(\mathbf{X} - \mathbf{v}\mathbf{v}^T\mathbf{X}) < \texttt{CLUSTER-TEST}(\mathbf{X}).$$

Where $\mathbf{v}$ is the first singular vector of $\mathbf{X}$. This leaves out two details: We need to whiten $\mathbf{X}$ and $\mathbf{X} - \mathbf{v}\mathbf{v}^T\mathbf{X}$, and we need to project $\mathbf{X}$ and $\mathbf{X} - \mathbf{v}\mathbf{v}^T\mathbf{X}$ to a common dimension. We perform this common projection for two reasons. 1) We use monte-carlo simulation to estimate $\mathbb{P}_{D,N}(C)$, and we would need to cache our $\mathbb{P}_{D,N}(C)$ for every possible D, potentially 1000s of them. By projecting to a common, smaller

Algorithm 1: INTERSECT-TEST

```
Input: $\mathbf{X}_{D\times K}, W \leq 8$
Output: true if $\mathbf{v}$, the first singular vector of $\mathbf{X}$ is a plausible intersection of data in $\mathbf{X}$.
Let $r$ = RANK($\mathbf{X}$).
Let $W'$ = MIN($r - 2, W$)
Let $\mathbf{U}, s, \mathbf{V}^T = SVD(\mathbf{X})$.   ▷ $\mathbf{V}^T$ is implicitly whitened.
Let $\alpha_1$ = CLUSTER-TEST($\mathbf{V}^T[:, : W']$)
Let $\alpha_2$ = CLUSTER-TEST($\mathbf{V}^T[:, 1 : W' + 1]$)   ▷ In null space of $\mathbf{v}$
RETURN $\alpha_2 < \alpha_1$
```

dimension before our test, we only need to cache $\mathbb{P}_{D,N}(C)$ for a handful of D. 2) By definition, $\mathbf{X}$ and $\mathbf{X} - \mathbf{v}\mathbf{v}^T\mathbf{X}$ are of different rank. Therefore, they would use different expected angle distributions, and we can no longer fairly compare the cramer-von-mises test results to see if $\mathbf{X} - \mathbf{v}\mathbf{v}^T\mathbf{X}$ is less clustered than $\mathbf{X}$.

The full algorithm for INTERSECT-TEST is in algorithm 1 $W \leq 8$ is an additional hyperparameter of the common dimensionality in which to project the data before comparison. Anecdotally, a larger W leads to wider trees.

3 Proofs of Theoretical Results

Theorem 3.1. *Any union of subspaces, or distribution lying on a union of subspaces can be equivalently represented by a PCT using the same or fewer parameters.*

Proof. We prove this result for a union / mixture of two multivatiate normal distributions which lie in a union of subspaces, but these results apply for any distribution embedded in a union of subspaces, where the data lying on each subspace can be described entirely by a covariance matrix.

Precise theorem: *In an even-prior mixture of two d-dimensional zero-mean Isotropic Gaussian's in a D-dimensional ambient space ($d < D$), where the intersection of the two gaussians is a c-dimensional subspace, we can efficiently and exactly represent the mixture with $2d - c$ vector-valued parameters corresponding to a PCT structure. Formally, let $p(\mathbf{x}) \sim \frac{1}{2}N(0, \mathbf{\Sigma}_A) + \frac{1}{2}N(0, \mathbf{\Sigma}_B)$ where $\Sigma_A = \mathbf{AIA}^T, \Sigma_B = \mathbf{BIB}^T$. This representation requires $2d$ vector-valued parameters: the columns of $\mathbf{A}$ and $\mathbf{B}$. If the intersection of $\mathbf{A}$ and $\mathbf{B}$ is a rank c subspace, then we can decompose $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_B$ into $\mathbf{\Sigma}_C + \mathbf{\Sigma}_{A\perp C}$ and $\mathbf{\Sigma}_C + \mathbf{\Sigma}_{B\perp C}$, respectively.*

Body of Proof. Let us take a singular value decomposition of $\mathbf{A}^T\mathbf{B} = \mathbf{W\Lambda M}^T$, where by definition of principle angles $\mathbf{\Lambda}$ is a diagonal matrix whose first c diagonals are 1, and the others are less than 1. Let $\mathbf{\Lambda}_S$ be the matrix with only a partial diagonal of c 1s, all other values 0.

Part 1: Decomposition of $\mathbf{\Sigma}_A = \mathbf{AA}^T$ and $\mathbf{\Sigma}_B = \mathbf{BB}^T$.

Let $\mathbf{P} = \mathbf{W\Lambda}_S\mathbf{W}^T$.

$$\begin{aligned}\mathbf{AA}^T &= \mathbf{APA}^T + \mathbf{A}(\mathbf{I} - \mathbf{P})\mathbf{A}^T \\ &= \mathbf{APA}^T + (\mathbf{A} - \mathbf{AP})\mathbf{A}^T \\ &= \mathbf{APA}^T + \mathbf{AA}^T - \mathbf{APA}^T \\ &= \mathbf{AA}^T\end{aligned}$$

(Similar for $\mathbf{B}$)

Let $\mathbf{U\Sigma V}^T$ be the SVD of $(\mathbf{AW\Lambda}_S)(\mathbf{BW\Lambda}_S)^T = \mathbf{AW\Lambda}_S\mathbf{W}^T\mathbf{B}^T$. Note the following: $V\mathbf{\Sigma U}^T = (\mathbf{BW\Lambda}_S)(\mathbf{AW\Lambda}_S)^T = \mathbf{BW\Lambda}_S\mathbf{W}^T\mathbf{A}^T$. $\mathbf{A}^T\mathbf{A} = \mathbf{B}^T\mathbf{B} = \mathbf{I}$, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{W}^T\mathbf{W} = \mathbf{I}$. $\mathbf{\Lambda}_S = \mathbf{\Lambda}_S^T = \mathbf{\Lambda}_S^2$. $\mathbf{\Sigma}$ also has a partial diagonal of c 1s, $\mathbf{\Sigma} = \mathbf{\Sigma}^T = \mathbf{\Sigma}^2$.

Part 2: $\mathbf{APA}^T = \mathbf{BPB}^T = \mathbf{U\Sigma U}^T$.

$$\begin{aligned}\mathbf{U\Sigma V}^T &= (\mathbf{AW\Lambda}_S)(\mathbf{BW\Lambda}_S)^T \\ \mathbf{U\Sigma V}^T\mathbf{BW\Lambda}_S &= (\mathbf{AW\Lambda}_S)(\mathbf{BW\Lambda}_S)^T\mathbf{BW\Lambda}_S \\ \mathbf{U\Sigma V}^T\mathbf{BW\Lambda}_S &= \mathbf{AW\Lambda}_S \\ \mathbf{U\Sigma V}^T\mathbf{BW\Lambda}_S\mathbf{W}^T\mathbf{A}^T &= \mathbf{AW\Lambda}_S\mathbf{W}^T\mathbf{A}^T \\ \mathbf{U\Sigma V}^T\mathbf{V\Sigma U}^T &= \mathbf{APA}^T \\ \mathbf{U\Sigma U}^T &= \mathbf{APA}^T\end{aligned}$$

(Similarly, for $\mathbf{V\Sigma V}^T = \mathbf{BPB}^T$)

Part 3: $\mathbf{APA}^T = \mathbf{BPB}^T = \mathbf{U\Sigma U}^T = \mathbf{V\Sigma V}^T$.

$$\begin{aligned}\mathbf{APA}^T &= \mathbf{BPB}^T \\ \mathbf{AW\Lambda}_S\mathbf{W}^T\mathbf{A}^T &= \mathbf{BW\Lambda}_S\mathbf{W}^T\mathbf{B}^T \\ \mathbf{U\Sigma V}^T\mathbf{BW\Lambda}_S\mathbf{W}^T\mathbf{A}^T &= \mathbf{V\Sigma U}^T\mathbf{AW\Lambda W}^T\mathbf{B}^T \\ \mathbf{AW\Lambda}_S\mathbf{W}^T\mathbf{B}^T\mathbf{V\Sigma U}^T &= \mathbf{BW\Lambda}_S\mathbf{W}^T\mathbf{A}^T\mathbf{U\Sigma V}^T \\ \mathbf{U\Sigma V}^T\mathbf{V\Sigma U}^T &= \mathbf{V\Sigma U}^T\mathbf{U\Sigma V}^T \\ \mathbf{APA}^T = \mathbf{U\Sigma U}^T &= \mathbf{V\Sigma V}^T = \mathbf{BPB}^T\end{aligned}$$

Going back to the original decomposition of the gaussian covariance matrices $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_B$ into $\mathbf{\Sigma}_C + \mathbf{\Sigma}_{A\perp C}$ and $\mathbf{\Sigma}_C + \mathbf{\Sigma}_{B\perp C}$, we have $\mathbf{\Sigma}_C = \mathbf{U\Sigma U}^T$. Because $\mathbf{\Sigma}$ only has c non-zero elements on it's diagonal, $\mathbf{\Sigma}_C$ can be represented by c vector-valued parameters. Furthermore $\mathbf{\Sigma}_{A\perp C} = \mathbf{A}(\mathbf{I} - \P)\mathbf{A}^T$, which needs only $d - c$ parameters. In total, the Gaussian mixture needs only $2d - c$ parameters, which is the number of nodes needed by the equivalent PCT. □

Theorem 3.2. *When presented data following a multivariate normal distribution, PCT construction will reduce to PCA, yielding a degenerate tree with probability at least $(1 - \alpha_{test})^D$. Formally, If $\mathbf{X} \in \mathbb{R}^D \sim N(0, \Sigma)$, then the constructed PCT tree will have D nodes and no splits with at least probability $1 - (\alpha_{test})^D$. In this case, the vector $\mathbf{v}_i$ of node $\mathcal{N}_i$ at height h_i will be equivalent to the h_ith singular vector of $\mathbf{X}$.*

Proof. If the data is truly multivariate normal, then any linear transformation of the data will also be multivariate normal, such as the linear transformation used to obtain the residual $\mathbf{R}_i = (\mathbf{I} - \mathbf{V}_i\mathbf{V}_i^T)\mathbf{X}$. Therefore, when we test for

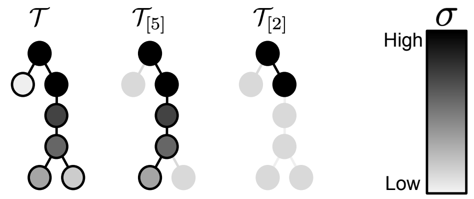


Figure 4: Illustration of optimal subtrees of a PCT.

subspace clustering in such data or any of it's node residuals using $\texttt{CLUSTER-TEST}(\mathbf{R_i})$, we will erroneously split a node with probability α_{test}. We repeat this test D times. Furthermore, if we never expand a node into two nodes, each subsequent node vector is a singular vector of $\mathbf{X}$. □

We also prove that our PCT learning algorithm leads to two simple but useful properties of PCTs. Firstly, nodes are orthogonal to their ancestors, making calculation of data approximations fast and decomposable. Secondly, node singular values decrease with height. Together these two properties are used in theorem 3, which provides a simple and efficient way to select the best approximating subtree of an existing tree by selecting its most significant nodes. This is like selecting the best principle components of PCA.

Property 3.1. *The singular vector of any node is orthogonal to those of its ancestors. Formally,*

$$\mathbf{v}_i \perp \mathbf{v}_j \;\; \forall j \in \mathcal{A}_i$$

as a consequence, all ancestral bases are orthonormal.

Proof. This property arises from the tree construction. Consider a node $\mathcal{N}_i$ and it's parent $\mathcal{P}_i$. $\mathbf{v}_i$ is either a singular vector of $\mathbf{R}_{\mathcal{P}_i}$ or a singular vector of a subset of $\mathbf{R}_{\mathcal{P}_i}$. Note that $\mathbf{R}_{\mathcal{P}_i} = \mathbf{X}_{\mathcal{P}_i} - \mathbf{V}_{\mathcal{P}_i}\mathbf{V}_{\mathcal{P}_i}^T\mathbf{X}_{\mathcal{P}_i}$. Therefore, each $\mathbf{v}_i$ is somewhere in the null space of $\mathbf{V}_{\mathcal{P}_i}$, its parent's basis, so $\mathbf{v}_i \perp \mathbf{v}_{\mathcal{P}_i}$. By induction, it is orthogonal to all its ancestors. □

Property 3.2. *A node's singular value is less than or equal to its ancestors. Formally,*

$$\sigma_i \leq \sigma_j \;\; \forall j \in \mathcal{A}_i.$$

Proof. We show this by contradiction. Consider a node $\mathcal{N}_i$, it's parent $\mathcal{P}_i$, and it's grandparent $\mathcal{P}_i^2$. Let $\overline{\mathbf{R}_{\mathcal{P}_i^2}}$ be the subset of $\mathbf{R}_{\mathcal{P}_i^2}$ by which the parent $\mathcal{N}_{\mathcal{P}_i}$ is constructed. Both $\mathbf{v}_{\mathcal{P}_\mathrm{i}}$ and $\mathbf{v}_\mathrm{i}$ are derived from $\overline{\mathbf{R}_{\mathcal{P}_i^2}}$, but $\mathbf{v}_{\mathcal{P}_\mathrm{i}}$ and $\sigma_{\mathcal{P}_i}$ are exactly it's first singular vector and value. If $\sigma_i > \sigma_{\mathcal{P}_i}$, it would violate the Eckart-Young theorem. Therefore, $\sigma_i \leq \sigma_{\mathcal{P}_i}$. By induction, $\sigma_i \leq \sigma_{j \in \mathcal{A}_i}$. □

Property 3.3. *We define the subtree of $\mathcal{T}$ consisting of the n most significant nodes as:*

$$\mathcal{T}_{[n]} = \{\mathcal{N}_i | \sigma_i \geq \sigma_{(n)}\}, \{\mathcal{E}_{ij} | \sigma_i, \sigma_j \geq \sigma_{(n)}\} \tag{1}$$

where $\sigma_{(n)}$ is the n_{th} largest singular value of nodes in $\mathcal{T}$. This is the best size n approximating subtree of $\mathcal{T}$. Formally,

$$\operatorname*{arg\,min}_{\mathcal{T}^* \subset \mathcal{T}} ||\mathbf{X} - \hat{\mathbf{X}}_{\mathcal{T}^*}||_F := \mathcal{T}_{[n]}$$

Note that we use a slightly non-traditional definition of subtree. In the literature, subtrees of $\mathcal{T}$ are usually defined as one of $\mathcal{T}$'s nodes and all it's descendents. We instead define a subtree of $\mathcal{T}$ as any subset of the nodes and edges of $\mathcal{T}$ which are connected. See Figure 4 for subtree examples.

Proof. We prove this in three parts. 1) We show that property 3.1 means that the approximation error is inversely proportional to the sum of singular values in the tree. 2) We can therefore minimize the approximation error by selecting the nodes with largest singular values. 3) Property 3.2 implies that $\mathcal{T}_{[n]}$ will share the root of $\mathcal{T}$, and will have no *gaps*, that is $\mathcal{N}_i \in \mathcal{T}_{[n]} \implies \mathcal{N}_j \in \mathcal{T}_{[n]} \;\forall\; \mathcal{N}_j \in \mathcal{A}_i$.

In order to prove this, we introduce *branches*, an additional organizing structure of Principle Component Trees. A branch $\mathcal{B}^b, b \in \{1...|\mathcal{B}|\}$ is the set of nodes that includes a single *leaf* node $\mathcal{N}_b$ and all its ancestors $\mathcal{A}_b$. $\mathbf{V}^b$ and $\mathbf{X}^b$ are simply the ancestral basis and assigned data for some leaf node $\mathcal{N}_b$. A single node can belong to one or more branches, and we denote the branches node $\mathcal{N}_i$ is a part of $\mathcal{B}_i$. Notationally, superscript = branch, subscript = node.

It is convenient for us to re-define our node approximations in terms of branches as follows:

$$\hat{\mathbf{X}} := \bigcup_{b=1}^{|\mathcal{B}|} \hat{\mathbf{X}}^b \qquad \hat{\mathbf{X}}^b := \mathbf{V}^b\mathbf{V}^{bT}\mathbf{X}^b$$

Furthermore, because all nodes in a branch are orthogonal, we can further decompose $\hat{\mathbf{X}}^b$ into the rank-1 approximations given by each node in the branch:

$$\hat{\mathbf{X}}^b := \mathbf{V}^b\mathbf{V}^{bT}\mathbf{X}^b := \sum_{i \in \mathcal{B}^b}^{|\mathcal{B}^b|} \mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b$$

Body of Proof

Part 1: Show that the quality of the tree approximation is given by the sum of the singular values of the nodes. We seek to minimize $||\mathbf{X}-\hat{\mathbf{X}}||_F$. Each element $\hat{\mathbf{x}}_k$ of $\hat{\mathbf{X}}$ is an orthonormal projection of the corresponding point $\mathbf{x}_k$. Via the Cauchy-Schwarz Inequality, we know that $||\hat{\mathbf{x}}_k|| < ||\mathbf{x}_k||$. As such, minimizing $||\mathbf{X} - \hat{\mathbf{X}}||_F$ is equivalent to maximizing $||\hat{\hat{\mathbf{X}}}||_F$. Furthermore, the sum of singular values $\sum_{i=1}^{|\mathcal{T}|} \sigma_i$ is equal to the norm of the approximation $||\hat{\mathbf{X}}||_F$.

$$||\hat{\mathbf{X}}||_F = \sum_{i=1}^{|\mathcal{T}|} \sigma_i \tag{1a}$$

$$||\bigcup_{b \in \mathcal{B}} \hat{\mathbf{X}}^b|| = \sum_{i=1}^{|\mathcal{T}|} ||\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}_i||_F \tag{1b}$$

$$||\bigcup_{b \in \mathcal{B}} \sum_{i=1}^{|\mathcal{B}^b|} \mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b|| = \sum_{i=1}^{|\mathcal{T}|}\sum_{b=1}^{|\mathcal{B}_i|} ||\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b||_F \tag{1c}$$

$$\sum_{b=1}^{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}^b|} ||\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b||_F = \sum_{i=1}^{|\mathcal{T}|}\sum_{b=1}^{|\mathcal{B}_i|} ||\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b||_F \tag{1d}$$

$$= \sum_{b=1}^{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{T}|} 1_{\{\mathcal{N}_i \in \mathcal{B}^b\}} ||\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^b||_F \tag{1e}$$

- 1b) Left: Decompose our full approximation in terms of single branch approximations. Right: Rephrase singular values as norm of approximations given by individual nodes onto their data.
- 1c) Left: Decompose each branch approximation into rank-1 approximations given by each node in the branch. Right: Decompose a node's data assignments as a union of the data assignments for each branch the node is in.
- 1d) Left: Norm of union of branch approximations → sum of norm of branch approximations. Recall that each branch's data approximations is disjoint.
- 1e) Reduce both sides to sum over all branches and all nodes, where the node $\mathcal{N}_i$ is in $\mathcal{B}^b$.

Part 2: To select the best size n approximating subtree of $\mathcal{T}$, select the top n nodes from $\mathcal{T}$ in terms of singular values. From part 1, we show that minimizing $||\mathbf{X} - \hat{\mathbf{X}}||_F$ is equivalent to maximizing $\sum_{i=1}^{|\mathcal{T}|} \sigma_i$. To do this maximization, we just select the largest singular value nodes from $\mathcal{T}$ n.

Part 3: These selected nodes form a contiguous subtree of $\mathcal{T}$. Property 3.2 proves that the singular values of the ancestors of a node are larger than that node's singular value. This implies that if a node $\mathcal{N}_i$ is chosen for the optimal approximating subtree, then all of it's ancestors $\mathcal{A}_i$ are also chosen for the subtree. It follows that the root of the original tree $\mathcal{T}$ is also the root of the subtree $\mathcal{T}_{[n]}$. Additionally, there are no "gaps" in the subtree; it remains connected, and the edges of the subtree $\mathcal{E}_{[n]}$ is a subset of the edges of original tree. The fact that this optimal subtree / subset of the PCT is connected and rooted at the most important node is analogous to the fact that the optimal subset of principle components in PCA is contiguous and includes the most important components by singular value. □

4 Experimental Results P-Values

.

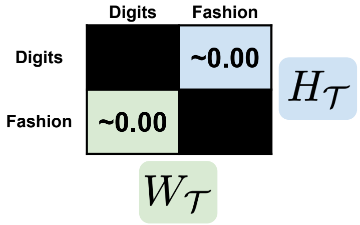
Within each of the three experiments in the main paper, we compare the PCT structure between two or more groups. Specifically, we present confidence intervals and scatter plots showing the distribution of our two summary statistics of tree structure: The *area under height curve* $H_{\mathcal{T}}$, and *area under width curve*, $W_{\mathcal{T}}$. These statistics are compared between bootstrap samples.

Here, we show the specific p-values indicating statistically significant differences in the subspace structure of our distributions. We compare $H_{\mathcal{T}}$ and $W_{\mathcal{T}}$ between distributions using two sample Wilcoxon Rank Sum Test of medians. See the following figures.

5 Additional Figures

References

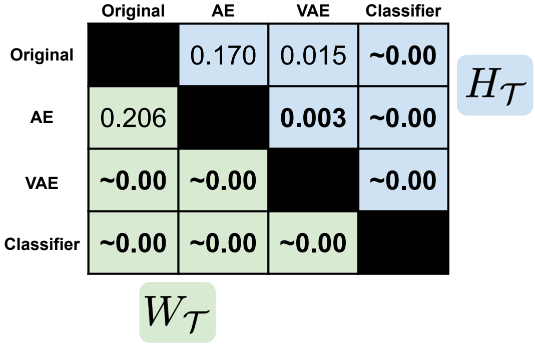
Thordsen, E.; and Schubert, E. 2022. ABID: Angle Based Intrinsic Dimensionality—Theory and analysis. *Information Systems*, 108: 101989.



	Digits	Fashion
Digits		**~0.00**
Fashion	**~0.00**	

$H_{\mathcal{T}}$
$W_{\mathcal{T}}$

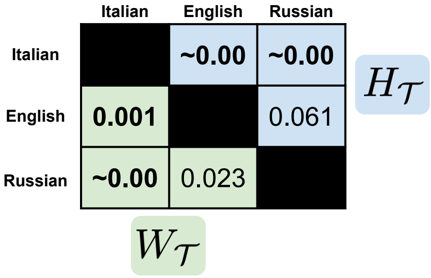
Figure 5: P-Values comparing the PCT structure of the MNIST Digits and MNIST Fashion datasets. Bold indicates statistically significant difference at the $\alpha = 0.01$ level.



	Original	AE	VAE	Classifier
Original		0.170	0.015	**~0.00**
AE	0.206		**0.003**	**~0.00**
VAE	**~0.00**	**~0.00**		**~0.00**
Classifier	**~0.00**	**~0.00**	**~0.00**	

$H_{\mathcal{T}}$
$W_{\mathcal{T}}$

Figure 6: P-Values comparing the PCT structure of Neural Network Latent Space. Bold indicates statistically significant difference at the $\alpha = 0.01$ level.



	Italian	English	Russian
Italian		**~0.00**	**~0.00**
English	**0.001**		0.061
Russian	**~0.00**	0.023	

$H_{\mathcal{T}}$
$W_{\mathcal{T}}$

Figure 7: P-Values comparing the PCT structure of word embeddings across languages. Bold indicates statistically significant difference at the $\alpha = 0.01$ level.

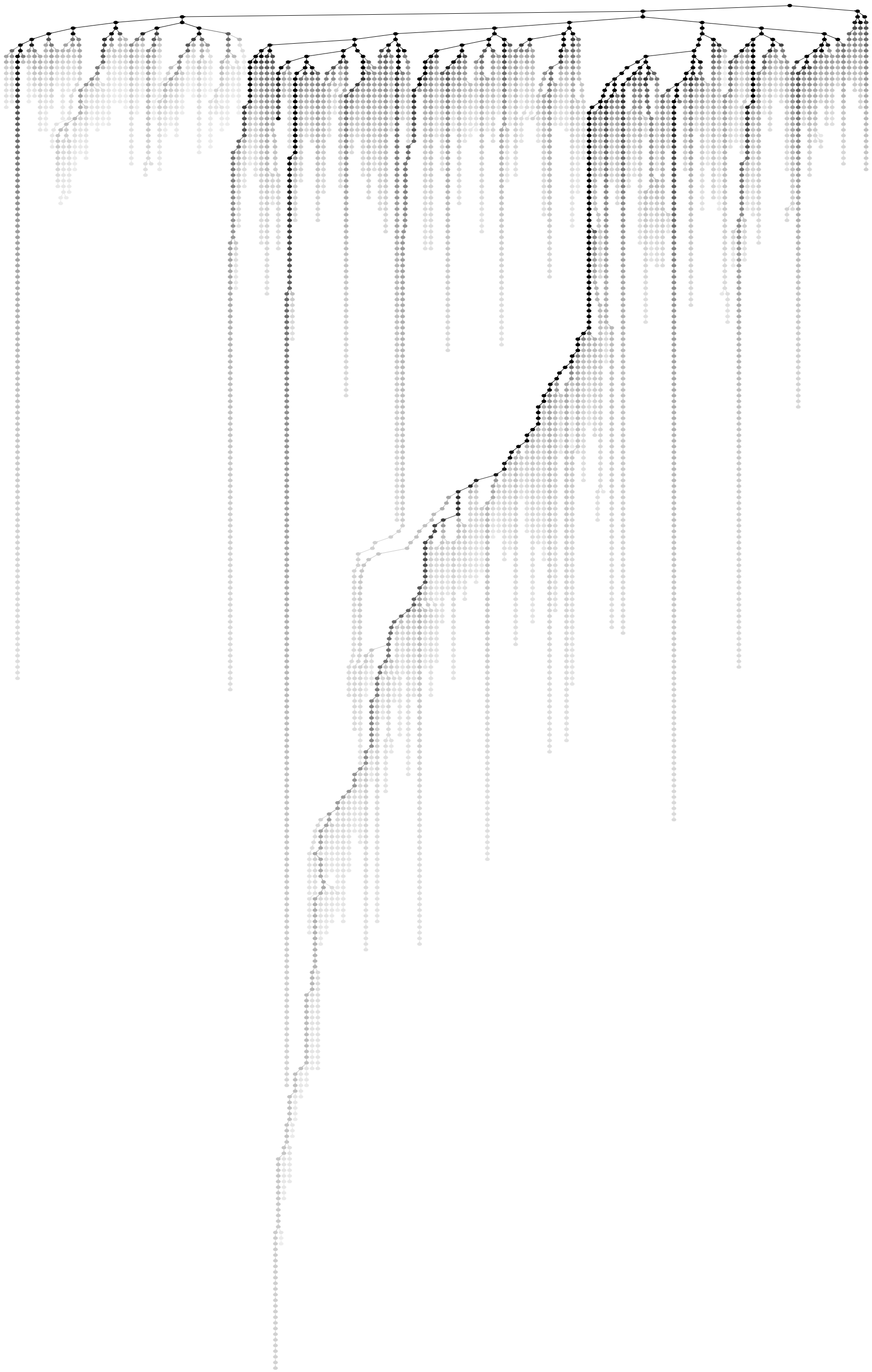


Figure 8: Full PCT for MNIST Digits. Singular Values Shaded.

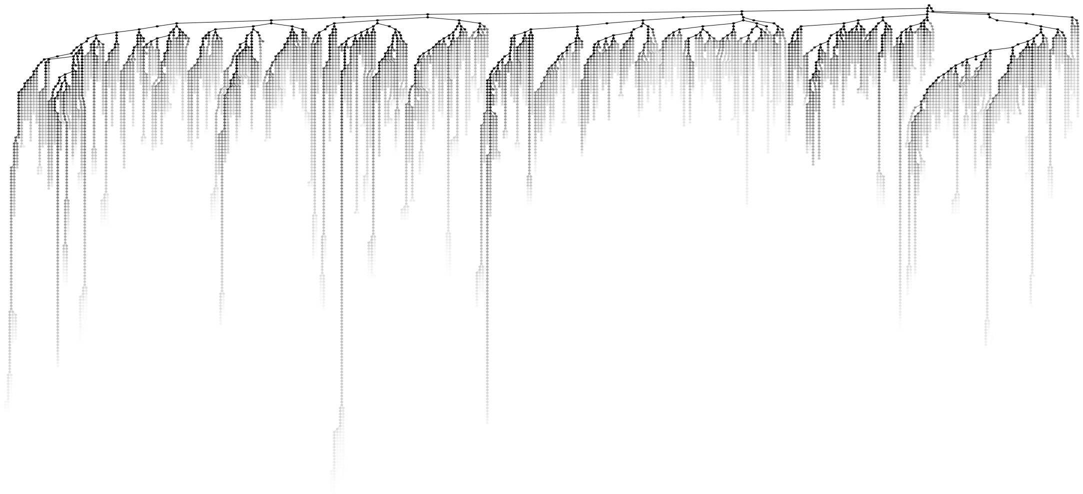


Figure 9: Full PCT for MNIST Fashion. Singular Values Shaded.