A Identifiability of latents drawn from a vMF around cluster vectors

In this section, we formally state and prove our core theoretical result. We start off by defining and discussing a useful notion, then introduce our assumptions on the data generating process. We proceed with the main statement and finish with the proof.

A.1 Affine Generator Systems

Definition 1 (Affine Generator System)**.** *A system of vectors $\{\boldsymbol{v}_c \in \mathbb{R}^d | c \in \mathscr{C}\}$ is called an* affine generator system *if the affine hull defined by them is $\mathbb{R}^d$. More precisely, any vector in $\mathbb{R}^d$ is an affine linear combination of the vectors in the system. Put into symbols: for any $\boldsymbol{v} \in \mathbb{R}^d$ there exist coefficients $\alpha_c \in \mathbb{R}$, such that*

$$\boldsymbol{v} = \sum_{c \in \mathscr{C}} \alpha_c \boldsymbol{v}_c \quad \textit{and} \quad \sum_{c \in \mathscr{C}} \alpha_c = 1. \tag{3}$$

Lemma 1 (Properties of affine generator systems)**.** *The following hold for any affine generator system $\{\boldsymbol{v}_c \in \mathbb{R}^d | c \in \mathscr{C}\}$:*

1. *for any $a \in \mathscr{C}$ the system $\{\boldsymbol{v}_c - \boldsymbol{v}_a | c \in \mathscr{C}\}$ is now a generator system of $\mathbb{R}^d$;*
2. *the invertible linear image of an affine generator system is also an affine generator system.*

A.2 Assumptions and main result

Assumptions 1C (DGP with vMF samples around cluster vectors)**.** *Assume the following DGP:*

(i) There exists a finite set of classes $\mathscr{C}$, represented by a set of unit-norm d-dimensional cluster-vectors $\{\boldsymbol{v}_c | c \in \mathscr{C}\} \subseteq \mathbb{S}^{d-1}$ such that they form an affine generator system of $\mathbb{R}^d$.

(ii) There is a finite set of instace labels $\mathscr{I}$ and a well-defined, surjective class function *$\mathcal{C} : \mathscr{I} \to \mathscr{C}$ (every label belongs to exactly one class and every class is in use).*

(iii) Our data sample is labelled with an instance label chosen uniformly, i.e., $I \in Uni(\mathscr{I})$ and, hence, belongs to class $C = \mathcal{C}(I)$.

(iv) The latent $\boldsymbol{z} \in \mathbb{S}^{d-1}$ of our data sample with label I is drawn from a vMF distribution around the cluster vector $\boldsymbol{v}_C$, where $C = \mathcal{C}(I)$:

$$\boldsymbol{z} \sim p(\boldsymbol{z}|C) \propto e^{\alpha \langle \boldsymbol{v}_C, \boldsymbol{z} \rangle}. \tag{4}$$

(v) The data sample $\boldsymbol{x}$ is generated by passing the latent $\boldsymbol{z}$ through a continuous and injective generator function $\boldsymbol{g} : \mathbb{S}^{d-1} \to \mathbb{R}^D$, i.e., $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z})$.

Assume that, using the DIET objective (6), we train a continuous encoder $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^d$ on $\boldsymbol{x}$ and a linear classification head $\boldsymbol{W}$ on top of $\boldsymbol{f}$. The rows of $\boldsymbol{W}$ are $\left\{\boldsymbol{w}_i^\top \mid i \in \mathscr{I}\right\}$. In other words, $\boldsymbol{W}$ computes similarities (scalar products) between its rows and the embeddings:

$$\boldsymbol{W} : \boldsymbol{f}(\boldsymbol{x}) \mapsto \left[\langle \boldsymbol{w}_i, \boldsymbol{f}(\boldsymbol{x}) \rangle \mid_{i \in \mathscr{I}}\right]. \tag{5}$$

In DIET, we optimize the following objective amongst all possible continuous encoders $\boldsymbol{f}$, linear classifiers $\boldsymbol{W}$, and $\beta > 0$:

$$\mathcal{L}(\boldsymbol{f}, \boldsymbol{W}, \beta) = \mathbb{E}_{(\boldsymbol{x}, I)}\left[-\ln \frac{e^{\beta \langle \boldsymbol{w}_I, \boldsymbol{f}(\boldsymbol{x}) \rangle}}{\sum_{j \in \mathscr{I}} e^{\beta \langle \boldsymbol{w}_j, \boldsymbol{f}(\boldsymbol{x}) \rangle}}\right] \tag{6}$$

Theorem 1C (Identifiability of latents drawn from a vMF around cluster vectors)**.** *Let $(\boldsymbol{f}, \boldsymbol{W}, \beta)$ globally minimize the DIET objective* (6) *under the following additional constraints:*

C1. both the embeddings $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{w}_i$'s are unit-normalized. Then:

(a) $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$ is orthogonal linear, i.e., the latents are identified up to an orthogonal linear transformation;

(b) $\boldsymbol{w}_i = \boldsymbol{h}(\boldsymbol{v}_{\mathcal{C}(i)})$ for any $i \in \mathscr{I}$, i.e., $\boldsymbol{w}_i$'s identify the cluster-vectors $\boldsymbol{v}_c$ up to the same orthogonal linear transformation;

(c) $\beta = \alpha$, the temperature of the vMF distribution is also identified.

C2. the embeddings $\boldsymbol{f}(\boldsymbol{x})$ are unit-normalized, the $\boldsymbol{w}_i$'s are unnormalized. Then:

(a) $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$ is orthogonal linear;

(b) $\boldsymbol{w}_i = \frac{\alpha}{\beta} \boldsymbol{h}(\boldsymbol{v}_{\mathcal{C}(i)}) + \boldsymbol{\psi}$ for any $i \in \mathscr{I}$, where $\boldsymbol{\psi}$ is a constant vector independent of i.

C3. the embeddings $\boldsymbol{f}(\boldsymbol{x})$ are unnormalized, while the $\boldsymbol{w}_i$'s are unit-normalized. If the system $\{\boldsymbol{v}_c|c\}$ ***is diverse enough in the sense of*** *Assum. 2, then:*
   (a) $\boldsymbol{w}_i = \mathcal{O}\boldsymbol{v}_{\mathcal{C}(i)}$, for any $i \in \mathscr{I}$, where $\mathcal{O}$ is orthogonal linear;
   (b) $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g} = \frac{\alpha}{\beta}\mathcal{O}$ with the same orthogonal linear transformation, but scaled with $\frac{\alpha}{\beta}$.

C4. neither the embeddings $\boldsymbol{f}(\boldsymbol{x})$ nor the rows of $\boldsymbol{W}$ are unit-normalized. Then:
   (a) $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$ is linear;
   (b) $\boldsymbol{w}_i$ identifies $\boldsymbol{v}_{\mathcal{C}(i)}$ up to an affine linear transformation.

Furthermore, in all cases, the row vectors that belong to samples of the same class are equal, i.e., for any $i, j \in \mathscr{I}$, $\mathcal{C}(i) = \mathcal{C}(j)$ implies $\boldsymbol{w}_i = \boldsymbol{w}_j$.

Remark. *In cases C2 and C4, the cluster vectors are unnormalized and, therefore, can absorb the temperature parameter β. Thus β can be set to 1 without loss of generality. In case C3, it is $\boldsymbol{f}$ that can absorb β.*

Assumption 2 (Diverse data)**.** *The system $\{\boldsymbol{v}_c|c \in \mathscr{C}\}$ is said to be diverse enough, if the following $|\mathscr{C}| \times 2d$ matrix has full column rank of $2d$:*

$$\begin{pmatrix} \cdots\cdots\cdots & \cdots\cdots\cdots \\ (\boldsymbol{v}_c \odot \boldsymbol{v}_c)^\top & \boldsymbol{v}_c^\top \\ \cdots\cdots\cdots & \cdots\cdots\cdots \end{pmatrix}, \tag{7}$$

where $[\boldsymbol{x} \odot \boldsymbol{y}]_i = x_i y_i$ is the elementwise- or Hadamard product.

As long as $|\mathscr{C}| \geq 2d$, this property holds almost surely w.r.t. the Lebesgue-measure of $\mathbb{S}^{d-1}$ or any continuous probability distribution of $\boldsymbol{v}_c \in \mathbb{S}^{d-1}$.

Proof. **Step 1: Deriving an equation characterizing the global optimizers of the objective.**

Rewriting the objective in terms of latents: we plug the expression $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z})$ into the optimization objective (6) to express the dependence in terms of the latents $\boldsymbol{z}$:

$$\mathcal{L}(\boldsymbol{f}, \boldsymbol{W}, \beta) = \mathbb{E}_{(\boldsymbol{z},I)}\left[-\ln \frac{e^{\beta\langle \boldsymbol{w}_I, \boldsymbol{f}\circ\boldsymbol{g}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{f}\circ\boldsymbol{g}(\boldsymbol{z})\rangle}}\right] = \mathcal{L}_{\boldsymbol{z}}(\boldsymbol{f}\circ\boldsymbol{g}, \boldsymbol{W}, \beta), \tag{8}$$

where the optimization is still over $\boldsymbol{f}$ (and not $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$).

We note that the generator $\boldsymbol{g}$ is, by assumption, continuously invertible on the *compact* set $\mathbb{S}^{d-1}$. Therefore, its image $\boldsymbol{g}(\mathbb{S}^{d-1})$ is compact, too, and its inverse $\boldsymbol{g}^{-1}$ is also continuous. By Tietze's extension theorem [Wikipedia, 2024b], $\boldsymbol{g}^{-1}$ can be continuously extended to a function $\boldsymbol{F} : \mathbb{R}^D \to \mathbb{S}^{d-1}$. Therefore, any continuous function $\boldsymbol{h} : \mathbb{S}^{d-1} \to \mathbb{R}^d$ can take the role of $\boldsymbol{f} \circ \boldsymbol{g}$ by substituting $\boldsymbol{f} = \boldsymbol{h} \circ \boldsymbol{F}$ continuous, since now $\boldsymbol{f} \circ \boldsymbol{g} = \boldsymbol{h} \circ (\boldsymbol{F} \circ \boldsymbol{g}) = \boldsymbol{h} \circ id_{\mathbb{S}^{d-1}} = \boldsymbol{h}$.

Hence, minimizing $\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{f}\circ\boldsymbol{g}, \boldsymbol{W}, \beta)$ (and by extension $\mathcal{L}(\boldsymbol{f}, \boldsymbol{W}, \beta)$) for continuous $\boldsymbol{f}$ equates to minimizing $\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{h}, \boldsymbol{W}, \beta)$ for continuous $\boldsymbol{h}$:

$$\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{h}, \boldsymbol{W}, \beta) = \mathbb{E}_{(\boldsymbol{z},I)}\left[-\ln \frac{e^{\beta\langle \boldsymbol{w}_I, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}}\right]. \tag{9}$$

Expressing the condition for global optimality of the objective: We rewrite the objective (9) by 1) using the indicator variable $\delta_{I=i}$ of the event $\{I = i\}$ and 2) applying the law of total expectation:

$$\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{h}, \boldsymbol{W}, \beta) = \mathbb{E}_{(\boldsymbol{z},I)}\left[-\sum_{i\in\mathscr{I}} \delta_{I=i} \ln \frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}}\right] \tag{10}$$

$$= \mathbb{E}_{\boldsymbol{z}}\left[\mathbb{E}_I\left[-\sum_{i\in\mathscr{I}} \delta_{I=i} \ln \frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}} \,\middle|\, \boldsymbol{z}\right]\right]. \tag{11}$$

Using the properties that $\mathbb{E}\left[A\, f(B)\middle|B\right] = \mathbb{E}\left[A\middle|B\right] f(B)$ and that $\mathbb{E}[\delta_{I=i}] = \mathbb{P}(I=i)$, we conclude that:

$$\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{h}, \boldsymbol{W}, \beta) = \mathbb{E}_{\boldsymbol{z}}\left[-\sum_{i\in\mathscr{I}} \mathbb{E}_I\left[\delta_{I=i} \ln \frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}} \,\middle|\, \boldsymbol{z}\right]\right] \tag{12}$$

$$= \mathbb{E}_{\boldsymbol{z}}\left[-\sum_{i\in\mathscr{I}} \mathbb{E}_I\left[\delta_{I=i}\middle|\boldsymbol{z}\right] \ln \frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}}\right] \tag{13}$$

$$= \mathbb{E}_{\boldsymbol{z}}\left[-\sum_{i\in\mathscr{I}} \mathbb{P}(I=i|\boldsymbol{z}) \ln \frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}}\right]. \tag{14}$$

By Gibbs' inequality [Wikipedia, 2024a], the cross-entropy inside the expectation is globally minimized if and only if

$$\frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}} = \mathbb{P}(I=i|\boldsymbol{z}), \quad \text{for any } i \in \mathscr{I}. \tag{15}$$

Moreover, the entire expectation is globally minimized if and only if the above equality (15) holds almost everywhere for $\boldsymbol{z} \in \mathbb{S}^{d-1}$.

Using that instance label I is uniformly distributed, or $\mathbb{P}(I=j) = \mathbb{P}(I=i)$, the likelihood of the sample being in class i can be expressed via Bayes' theorem as:

$$\mathbb{P}(I=i|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|I=i)\mathbb{P}(I=i)}{\sum_{j\in\mathscr{I}} p(\boldsymbol{z}|I=j)\mathbb{P}(I=j)} = \frac{p(\boldsymbol{z}|I=i)}{\sum_{j\in\mathscr{I}} p(\boldsymbol{z}|I=j)}. \tag{16}$$

Substituting (16) into (15) yields that for any $i \in \mathscr{I}$ and almost everywhere w.r.t. $\boldsymbol{z} \in \mathbb{S}^{d-1}$:

$$\frac{e^{\beta\langle \boldsymbol{w}_i, \boldsymbol{h}(\boldsymbol{z})\rangle}}{\sum_{j\in\mathscr{I}} e^{\beta\langle \boldsymbol{w}_j, \boldsymbol{h}(\boldsymbol{z})\rangle}} = \frac{p(\boldsymbol{z}|I=i)}{\sum_{j\in\mathscr{I}} p(\boldsymbol{z}|I=j)}. \tag{17}$$

We now divide the equation (17) for the probability of a sample having label i with that of having label k and take the logarithm. This yields that $\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{h}, \boldsymbol{W}, \beta)$ is globally minimized if and only if

$$\beta\langle \boldsymbol{w}_i - \boldsymbol{w}_k, \boldsymbol{h}(\boldsymbol{z})\rangle = \ln \frac{p(\boldsymbol{z}|I=i)}{p(\boldsymbol{z}|I=k)} \tag{18}$$

holds for any $i, k \in \mathscr{I}$ and almost everywhere w.r.t. $\boldsymbol{z} \in \mathbb{S}^{d-1}$.

Plugging in the vMF distribution: Plugging the assumed conditional distribution from (4) into (18) yields the equivalent expression:

$$\beta\langle \boldsymbol{w}_i - \boldsymbol{w}_k, \boldsymbol{h}(\boldsymbol{z})\rangle = \alpha\langle \boldsymbol{v}_{\mathcal{C}(i)} - \boldsymbol{v}_{\mathcal{C}(k)}, \boldsymbol{z}\rangle \tag{19}$$

holds for any $i, k \in \mathscr{I}$ and almost everywhere w.r.t. $\boldsymbol{z} \in \mathbb{S}^{d-1}$. Since $\boldsymbol{h}$ is continuous, the equation holds almost everywhere w.r.t. $\boldsymbol{z}$ if and only if it holds for all $\boldsymbol{z} \in \mathbb{S}^{d-1}$.

Observe that if $\boldsymbol{h} = id|_{\mathbb{S}^{d-1}}$, $\boldsymbol{w}_i = \boldsymbol{v}_{\mathcal{C}(i)}$ for any $i \in \mathscr{I}$, and $\beta = \alpha$, then the equation is satisfied. Thus, we can conclude that the global minimum of the cross-entropy loss is achieved.

Step 2: Solving the equation for $\boldsymbol{h}, \boldsymbol{W}$ and proving identifiability.

We now find all solutions to prove the identifiability of the latent variables and that of the cluster vectors. Denote $\tilde{\boldsymbol{w}}_i = \frac{\beta}{\alpha}\boldsymbol{w}_i$ to simplify the above equation to:

$$\langle \tilde{\boldsymbol{w}}_i - \tilde{\boldsymbol{w}}_k, \boldsymbol{h}(\boldsymbol{z})\rangle = \langle \boldsymbol{v}_{\mathcal{C}(i)} - \boldsymbol{v}_{\mathcal{C}(k)}, \boldsymbol{z}\rangle. \tag{20}$$

$\boldsymbol{h}$ is injective and has full-dimensional image: We prove that $\boldsymbol{h}$ is injective. Assume that $\boldsymbol{h}(\boldsymbol{z}_1) = \boldsymbol{h}(\boldsymbol{z}_2)$ for some $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{S}^{d-1}$. Plugging $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ into (20) and subtracting the two equations yields:

$$0 = \langle \tilde{\boldsymbol{w}}_i - \tilde{\boldsymbol{w}}_k, \boldsymbol{h}(\boldsymbol{z}_1) - \boldsymbol{h}(\boldsymbol{z}_2)\rangle = \langle \boldsymbol{v}_{\mathcal{C}(i)} - \boldsymbol{v}_{\mathcal{C}(k)}, \boldsymbol{z}_1 - \boldsymbol{z}_2\rangle, \tag{21}$$

for any i, k. However, as the cluster vectors $\{\boldsymbol{v}_c|c\}$ form an affine generator system, the vectors $\{\boldsymbol{v}_{\mathcal{C}(i)} - \boldsymbol{v}_{\mathcal{C}(k)}|i,k\}$ form a generator system of $\mathbb{R}^d$ (see Lem. 1). Therefore, $\langle \boldsymbol{y}, \boldsymbol{z}_1 - \boldsymbol{z}_2\rangle = 0$, for any $\boldsymbol{y} \in \mathbb{R}^d$, which holds if and only if $\boldsymbol{z}_1 = \boldsymbol{z}_2$. Hence, $\boldsymbol{h}$ is injective.

By the Borsuk-Ulam theorem, for any continuous map from $\mathbb{S}^{d-1}$ to a space of dimensionality at most $d-1$ there exists some pair of antipodal points that are mapped to the same point. Consequently, no such function can be injective at the same time. Since $h : \mathbb{S}^{d-1} \to \mathbb{R}^d$ is injective, the linear span of its image must be $\mathbb{R}^d$.

Collapse of $\boldsymbol{w}_i$'s: We prove that $\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{w}}_k$ if $\mathcal{C}(i) = \mathcal{C}(k)$, i.e., samples from the same cluster will have equal rows of $\boldsymbol{W}$ associated with them.

Assume that $\mathcal{C}(i) = \mathcal{C}(k)$ and substitute them into (20):

$$\langle \tilde{\boldsymbol{w}}_i - \tilde{\boldsymbol{w}}_k, \boldsymbol{h}(\boldsymbol{z}) \rangle = 0 \quad \text{for any } \boldsymbol{z} \in \mathbb{S}^{d-1}. \tag{22}$$

However, we have just seen that the linear span of the image of $\boldsymbol{h}$ is $\mathbb{R}^d$, which implies that $\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{w}}_k$. Consequently, we may abuse out notation by setting $\tilde{\boldsymbol{w}}_c = \tilde{\boldsymbol{w}}_i$ if $\mathcal{C}(i) = c$, which yields a new form for (20):

$$\langle \tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b, \boldsymbol{h}(\boldsymbol{z}) \rangle = \langle \boldsymbol{v}_a - \boldsymbol{v}_b, \boldsymbol{z} \rangle, \tag{23}$$

for any $a, b \in \mathscr{C}$ and any $\boldsymbol{z} \in \mathbb{S}^{d-1}$.

Linear transformation from $\boldsymbol{v}_a - \boldsymbol{v}_b$ to $\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b$: We now prove the existence of a linear map $\mathcal{A}$ on $\mathbb{R}^d$ such that $\mathcal{A}(\boldsymbol{v}_a - \boldsymbol{v}_b) = \tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b$ for any $a, b \in \mathscr{C}$. For this, we prove that the following mapping is well-defined:

$$\mathcal{A} : \sum_{a,b \in \mathscr{C}} \lambda_{ab} (\boldsymbol{v}_a - \boldsymbol{v}_b) \mapsto \sum_{a,b \in \mathscr{C}} \lambda_{ab} (\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b). \tag{24}$$

Since the system $\{\boldsymbol{v}_a - \boldsymbol{v}_b | a, b\}$ is not necessarily linearly independent, we have to prove that the mapping is independent of the choice of the linear combination. More precisely if for some coefficients $\lambda_{ab}, \lambda'_{ab}$

$$\sum_{a,b \in \mathscr{C}} \lambda_{ab} (\boldsymbol{v}_a - \boldsymbol{v}_b) = \sum_{a,b \in \mathscr{C}} \lambda'_{ab} (\boldsymbol{v}_a - \boldsymbol{v}_b) \tag{25}$$

holds, then it should be implied that

$$\sum_{a,b \in \mathscr{C}} \lambda_{ab} (\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b) = \sum_{a,b \in \mathscr{C}} \lambda'_{ab} (\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b). \tag{26}$$

Assume that (25) holds. Then, the difference of the two sides is:

$$0 = \sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) (\boldsymbol{v}_a - \boldsymbol{v}_b). \tag{27}$$

Taking the scalar product with an arbitrary $\boldsymbol{z} \in \mathbb{S}^{d-1}$ and using the linearity of the scalar product gives us:

$$0 = \langle \sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) (\boldsymbol{v}_a - \boldsymbol{v}_b), \boldsymbol{z} \rangle = \sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) \langle \boldsymbol{v}_a - \boldsymbol{v}_b, \boldsymbol{z} \rangle. \tag{28}$$

Now using (23) yields:

$$0 = \sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) \langle \tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b, \boldsymbol{h}(\boldsymbol{z}) \rangle = \langle \sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) (\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b), \boldsymbol{h}(\boldsymbol{z}) \rangle. \tag{29}$$

However, the linear span of the image of $\boldsymbol{h}$ is $\mathbb{R}^d$, which implies that

$$\sum_{a,b \in \mathscr{C}} (\lambda_{ab} - \lambda'_{ab}) (\tilde{\boldsymbol{w}}_a - \tilde{\boldsymbol{w}}_b) = 0, \tag{30}$$

equivalent to (26). Therefore, the mapping is well-defined. The linearity of $\mathcal{A}$ follows trivially.

$\boldsymbol{h}$ is linear: Equation (23) becomes:

$$\langle \mathcal{A}(\boldsymbol{v}_a - \boldsymbol{v}_b), \boldsymbol{h}(\boldsymbol{z}) \rangle = \langle \boldsymbol{v}_a - \boldsymbol{v}_b, \boldsymbol{z} \rangle, \tag{31}$$

for any $a, b \in \mathscr{C}$ and any $\boldsymbol{z} \in \mathbb{S}^{d-1}$. Nevertheless, $\{\boldsymbol{v}_a - \boldsymbol{v}_b | a, b \in \mathscr{C}\}$ is a generator system of $\mathbb{R}^d$, and, hence, (31) is equivalent to

$$\langle \mathcal{A}\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{z}) \rangle = \langle \boldsymbol{y}, \boldsymbol{z} \rangle, \quad \text{for any } \boldsymbol{y} \in \mathbb{R}^d \text{ and any } \boldsymbol{z} \in \mathbb{S}^{d-1}. \tag{32}$$

This is further equivalent to

$$\langle \boldsymbol{y}, \mathcal{A}^\top \boldsymbol{h}(\boldsymbol{z}) \rangle = \langle \boldsymbol{y}, \boldsymbol{z} \rangle. \tag{33}$$

Since $\boldsymbol{y}$ is arbitrary, we conclude that $\mathcal{A}^\top \boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{z}$ for any $\boldsymbol{z} \in \mathbb{S}^{d-1}$. Therefore $\mathcal{A}$ is an invertible transformation and $\boldsymbol{h} = (\mathcal{A}^\top)^{-1}$ is linear.

Proving Thm. 1C case C4: We have shown that $\boldsymbol{h}$ is linear. Furthermore, from (31) it follows, by fixing b and defining $\boldsymbol{\psi} = \mathcal{A}\boldsymbol{v}_b - \boldsymbol{w}_b$, that

$$\tilde{\boldsymbol{w}}_a = \mathcal{A}\boldsymbol{v}_a + \boldsymbol{\psi}, \quad \text{for any } a \in \mathscr{C}, \tag{34}$$

which proves case C4 of Thm. 1C.

Proving Thm. 1C case C2: As a special case of the previous one, now we assume that $\boldsymbol{h}(\boldsymbol{z})$ is unit-normalized and maps $\mathbb{S}^{d-1}$ to $\mathbb{S}^{d-1}$. That amounts to $\boldsymbol{h} = (\mathcal{A}^\top)^{-1}$ being linear, norm-preserving, and therefore orthogonal. Consequently $\mathcal{A}$ is also orthogonal, $\boldsymbol{h} = \mathcal{A}$ and (34) simplifies to $\frac{\beta}{\alpha}\boldsymbol{w}_a = \tilde{\boldsymbol{w}}_a = \mathcal{A}\boldsymbol{v}_a + \boldsymbol{\psi} = \boldsymbol{h}(\boldsymbol{v}_a) + \boldsymbol{\psi}$, which proves C2 of Thm. 1C.

Proving Thm. 1C case C1: We now assume that both $\boldsymbol{h}$ and $\boldsymbol{w}_i$'s are unit-normalized. Consequently, $\boldsymbol{h} = \mathcal{A}$ is orthogonal linear and $\boldsymbol{w}_a = \frac{\alpha}{\beta}\mathcal{A}\boldsymbol{v}_a + \boldsymbol{\psi}$.

Therefore, on one hand, the $\boldsymbol{w}_a$'s lie on a d-dimensional hypersphere of radius $\frac{\alpha}{\beta}$ and center $\boldsymbol{\psi}$. On the other hand, by definition, $\boldsymbol{w}_a$'s also lie on the unit hypersphere $\mathbb{S}^{d-1}$.

Since the system $\{\boldsymbol{w}_a | a \in \mathscr{C}\}$ is the bijective affine linear image of the affine generator system $\{\boldsymbol{v}_a | a \in \mathscr{C}\}$, $\{\boldsymbol{w}_a | a \in \mathscr{C}\}$ is also an affine generator system (Lem. 1). Consequently, there could be at most one hypersphere in $\mathbb{R}^d$ which contains all the $\boldsymbol{w}_a$'s. Hence $\frac{\alpha}{\beta} = 1$, $\boldsymbol{\psi} = \boldsymbol{0}$, and $\boldsymbol{w}_a = \boldsymbol{h}(\boldsymbol{v}_a)$, which proves C1 of Thm. 1C.

Proving Thm. 1C case C3: Finally, we assume that $\boldsymbol{w}_i$'s are unit-normalized. As this is a special case of Thm. 1C C4, we know that there exists a constant vector $\boldsymbol{\psi}$ such that:

$$\boldsymbol{w}_a = \frac{\alpha}{\beta}\mathcal{A}\boldsymbol{v}_a + \boldsymbol{\psi}, \tag{35}$$

for any $a \in \mathscr{C}$. We are going to prove that $\mathcal{O} = \frac{\alpha}{\beta}\mathcal{A}$ is orthogonal and $\boldsymbol{\psi} = \boldsymbol{0}$.

Let $\mathcal{O} = \mathcal{U}^\top \Sigma \mathcal{V}$ be the singular value decomposition (SVD) of $\mathcal{O}$. Consequently, after premultiplying with $\mathcal{U}$, we receive:

$$\mathcal{U}\boldsymbol{w}_a = \Sigma\mathcal{V}\boldsymbol{v}_a + \mathcal{U}\boldsymbol{\psi}. \tag{36}$$

As orthogonal transformations $\mathcal{U}$ and $\mathcal{V}$ keep their arguments unit-normalized and $\{\mathcal{V}\boldsymbol{v}_a - \mathcal{V}\boldsymbol{v}_b\}$ is still an affine generator system (Lem. 1), we may assume without the loss of generality that

$$\boldsymbol{w}_a = \Sigma\boldsymbol{v}_a + \boldsymbol{\psi}, \tag{37}$$

for any $a \in \mathscr{C}$, where all $\boldsymbol{v}_a$'s and $\boldsymbol{w}_a$'s are unit-normalized.

Let us assume that $\boldsymbol{\psi} \neq \boldsymbol{0}$. In that case both sides of (37) can be scaled such that the offset $\boldsymbol{\psi}$ has unit norm. In this case $\boldsymbol{w}_a$'s are no longer on the unit hypersphere, but they instead have a mutual norm r. Assuming that the diagonal elements of Σ are $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$, this is equivalent to:

$$r^2 = \|\Sigma\boldsymbol{v}_a + \boldsymbol{\psi}\|^2 = \|\Sigma\boldsymbol{v}_a\|^2 + 2\langle\Sigma\boldsymbol{v}_a, \boldsymbol{\psi}\rangle + \|\boldsymbol{\psi}\|^2 \tag{38}$$

$$= \langle\boldsymbol{v}_a \odot \boldsymbol{v}_a, \boldsymbol{\sigma} \odot \boldsymbol{\sigma}\rangle + \langle\boldsymbol{v}_a, 2\boldsymbol{\sigma} \odot \boldsymbol{\psi}\rangle + 1, \tag{39}$$

where $[\boldsymbol{x} \odot \boldsymbol{y}]_i = x_i y_i$ is the elementwise product. Eq. (39) is equivalent to the following:

$$(\boldsymbol{v}_a \odot \boldsymbol{v}_a)^\top(\boldsymbol{\sigma} \odot \boldsymbol{\sigma}) + \boldsymbol{v}_a^\top(2\boldsymbol{\sigma} \odot \boldsymbol{\psi}) - r^2 = -1. \tag{40}$$

Collecting the equations for all $a \in \mathscr{C}$ yields:

$$\mathcal{D}\begin{pmatrix} \boldsymbol{\sigma} \odot \boldsymbol{\sigma} \\ 2\boldsymbol{\sigma} \odot \boldsymbol{\psi} \\ r^2 \end{pmatrix} = -\boldsymbol{1}_{|\mathscr{C}|}, \tag{41}$$

where $\mathcal{D}$ is the following $|\mathscr{C}| \times (2d+1)$ matrix:

$$\mathcal{D} = \begin{pmatrix} \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots \\ (\boldsymbol{v}_a \odot \boldsymbol{v}_a)^\top & \boldsymbol{v}_a^\top & -1 \\ \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots \end{pmatrix}. \tag{42}$$

By Assum. 2, the left $|\mathscr{C}| \times 2d$ submatrix of $\mathcal{D}$ has full rank of $2d$. Consequently, the solution space to the more general, linear equation $\mathcal{D}\boldsymbol{t} = -\boldsymbol{1}_{|\mathscr{C}|}$, where $\boldsymbol{t} \in \mathbb{R}^d$, has a dimensionality of at most 1.

Using the unit-normality of $\boldsymbol{v}_a$'s, we see that $(\boldsymbol{v}_a \odot \boldsymbol{v}_a)^\top \mathbf{1}_d = 1$. From this, it follows that the solutions are exactly the following:

$$\boldsymbol{t} = \begin{pmatrix} \gamma \cdot \mathbf{1}_d \\ \mathbf{0}_d \\ \gamma + 1 \end{pmatrix}, \quad \text{where } \gamma \in \mathbb{R}. \tag{43}$$

Therefore, for any solution of (41) there exists γ such that:

$$\boldsymbol{\sigma} \odot \boldsymbol{\sigma} = \gamma \cdot \mathbf{1}_d \tag{44}$$

$$\boldsymbol{\sigma} \odot \boldsymbol{\psi} = \mathbf{0}_d. \tag{45}$$

However, as the original transformation $\mathcal{A}$ was invertible, all singular values σ_i are strictly positive and, thus, it follows that $\boldsymbol{\psi} = \mathbf{0}$. Technically speaking, this is a contradiction to our initial assumption that $\boldsymbol{\psi} \neq \mathbf{0}$. All in all, it follows that $\boldsymbol{\psi} = \mathbf{0}$ is the only possibility.

Therefore, (37) becomes:

$$\boldsymbol{w}_a = \Sigma \boldsymbol{v}_a, \tag{46}$$

where all $\boldsymbol{v}_a$'s and $\boldsymbol{w}_a$'s are unit-normalized. Following the same derivation yields:

$$1 = \|\Sigma \boldsymbol{v}_a\|^2 = (\boldsymbol{v}_a \odot \boldsymbol{v}_a)^\top (\boldsymbol{\sigma} \odot \boldsymbol{\sigma}), \tag{47}$$

or, after collecting the equations for all $a \in \mathscr{C}$:

$$\mathcal{B}(\boldsymbol{\sigma} \odot \boldsymbol{\sigma}) = \mathbf{1}_{|\mathscr{C}|}, \tag{48}$$

where $\mathcal{B}$ is the $|\mathscr{C}| \times d$ matrix

$$\mathcal{B} = \begin{pmatrix} \cdots\cdots\cdots \\ (\boldsymbol{v}_a \odot \boldsymbol{v}_a)^\top \\ \cdots\cdots\cdots \end{pmatrix}. \tag{49}$$

By Assum. 2, $\mathcal{B}$ has full rank, thus, there is at most one solution to the equation $\mathcal{B}\boldsymbol{t} = \mathbf{1}_{|\mathscr{C}|}$. Due to the unit-normality of $\boldsymbol{v}_a$'s, this solution is exactly $\boldsymbol{t} = \mathbf{1}_d$. However, as the singular values σ_i are all positive, the only solution to $\boldsymbol{\sigma} \odot \boldsymbol{\sigma} = \mathbf{1}_d$ is $\boldsymbol{\sigma} = \mathbf{1}_d$. This is equivalent to saying that $\mathcal{O} = \frac{\alpha}{\beta}\mathcal{A}$ is orthogonal.

Furthermore, $\boldsymbol{h} = (\mathcal{A}^\top)^{-1} = (\frac{\beta}{\alpha}\mathcal{O}^\top)^{-1} = \frac{\alpha}{\beta}\mathcal{O}$.

□

B Additional experimental results

In Tab. 2, we present additional ablation studies exploring the effect of varying the levels of concentration for $\boldsymbol{v}_c$ across the unit hyper-sphere. We do not observe any significant impact on the R^2 scores from more concentrated cluster centroids $\boldsymbol{v}_c$.

Table 2: Identifiability in the synthetic setup. Mean ± standard deviation across 5 random seeds. Settings that match our theoretical assumptions are ✓. We report the R^2 score for linear mappings, $\tilde{\boldsymbol{z}} \to \boldsymbol{z}$ and $\boldsymbol{w}_i \to \boldsymbol{v}_c$ for cases with normalized (o) and unnormalized (a) $\boldsymbol{w}_i$. For unnormalized $\boldsymbol{w}_i$, we verify that mappings $\tilde{\boldsymbol{z}} \to \boldsymbol{z}$ are orthogonal by reporting the mean absolute error between their singular values and those of an orthogonal transformation.

						normalized $\boldsymbol{w}_i$ cases				unnormalized $\boldsymbol{w}_i$	
						$R^2_{\text{o}}(\uparrow)$		MAE$_{\text{o}}(\downarrow)$		$R^2_{\text{a}}(\uparrow)$	
N	d	$\lvert\mathscr{C}\rvert$	$p(\boldsymbol{v}_c)$	$p(\boldsymbol{z}\lvert\boldsymbol{v}_c)$	M.	$\tilde{\boldsymbol{z}} \to \boldsymbol{z}$	$\boldsymbol{w}_i \to \boldsymbol{v}_c$	$\tilde{\boldsymbol{z}} \to \boldsymbol{z}$	$\boldsymbol{w}_i \to \boldsymbol{v}_c$	$\tilde{\boldsymbol{z}} \to \boldsymbol{z}$	$\boldsymbol{w}_i \to \boldsymbol{v}_c$
10^3	5	100	Uniform	vMF(κ=10)	✓	$98.6_{\pm 0.01}$	$99.9_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.00_{\pm 0.00}$	$99.0_{\pm 0.00}$	$99.9_{\pm 0.00}$
10^3	5	100	Laplace	vMF(κ=10)	✓	$98.7_{\pm 0.00}$	$99.5_{\pm 0.00}$	$0.01_{\pm 0.00}$	$0.00_{\pm 0.00}$	$99.1_{\pm 0.00}$	$99.8_{\pm 0.00}$
10^3	5	100	Normal	vMF(κ=10)	✓	$98.2_{\pm 0.01}$	$99.2_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.00_{\pm 0.00}$	$99.2_{\pm 0.00}$	$99.8_{\pm 0.00}$

C Acronyms

CL Contrastive Learning

DGP data generating process

ICA Independent Component Analysis

LVM latent variable model

SSL Self-Supervised Learning

vMF von Mises-Fisher