HyperCom: Community Alignment across Social Networks in Hyperbolic Space Supplementary Materials

Anonymous Authors

A. Lemma (Positive Definiteness)

Proof. To proof the positive definiteness of $\mathbf{\Sigma}$, we first introduce an auxiliary matrix $\tilde{\mathbf{\Sigma}}$ defined as follows:

$$\tilde{\mathbf{\Sigma}} = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig} b_{ig} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g - \frac{\hat{\boldsymbol{\beta}}_g}{b_{ig}}\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g - \frac{\hat{\boldsymbol{\beta}}_g}{b_{ig}}\right)'.$$

Indeed, $a_{ig} = \mathbb{E}\left[W_{ig}\right]$ and $b_{ig} = \mathbb{E}\left[1/W_{ig}\right]$ with the formulation $\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}$, where $W \sim \mathcal{I}(\omega, 1, \lambda)$ and $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ are latent random variables (Browne and McNicholas 2015). Recall Jensen's inequality. We have $1/\mathbb{E}\left[W_{ig}\right] \leq \mathbb{E}\left[1/W_{ig}\right]$ for all i, i.e., $1/a_{ig} \leq b_{ig}$, and thus

$$\overline{a}_g = \frac{1}{n_g}\sum_{i=1} \hat{z}_{ig} a_{ig} \geq \frac{1}{n_g}\sum_{i=1}\frac{\hat{z}_{ig}}{b_{ig}}.$$

Finally, we have the following hold:

$$\mathbf{\Sigma} = \tilde{\mathbf{\Sigma}} + (\overline{a}_g - \frac{1}{n_g}\sum_{i=1}\frac{\hat{z}_{ig}}{b_{ig}})\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^T.$$

With the positive definiteness of $\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^T$, we ensure the positive definiteness of $\mathbf{\Sigma}$, or formally, we obtain an inequality $\hat{\mathbf{\Sigma}}_g \succeq \mathbf{\Sigma}_g^* \succeq 0$. □

B. On Identifiability

In the context of finite mixture, *identifiability* is of fundamental important as it allows for consistent estimation and data recovery. First, we review the concept of identifiability of a finite mixture $f(\mathbf{x}|\boldsymbol{\phi}) = \sum_{i=1}^{G} \pi_i f_i\left(\mathbf{x}|\boldsymbol{\phi}_i\right)$. This finite mixture is identifiable if distinct mixing $f_i\left(\mathbf{x}|\boldsymbol{\phi}_g\right)$ with finite mixing proportion $\{\pi_i\}_{i=1}^{G}$ corresponds to distinct mixtures (Holzmann, Munk, and Gneiting 2006). Formally, we give the definition as follows:

Definition (Identifiability of Finite Mixture). *Given a normal mean-variance family $f_i(\boldsymbol{x}|\phi_i)$ of d-dimensionality with parameter space of $\mathcal{S}^d$, the finite mixtures are identifiable if and only if*

$$\sum_{i=1}^{G} \pi_i f_i(\boldsymbol{x}|\phi_i) = \sum_{i=1}^{G} \pi'_i f_{i'}(\boldsymbol{x}|\phi'_i),$$

where $\boldsymbol{x} \in \mathbb{R}^d$, G is the given finite number of components in the mixture and $\sum_{i=1}^{G} \pi_i = \sum_{i=1}^{G} \pi'_i = 1$ with nonnegative π_i and π'_i for all $i = 1, 2, ..., G$, under the observations of $\boldsymbol{x}$, implies there exists a permutation $p(\cdot)$ such that for all i,

$$(\pi_i, \phi_i) = \left(\pi_{p(i)}, \phi_{p(i)}\right).$$

Evidently, $f(\mathbf{x}|\boldsymbol{\phi})$ is said to be identifiable if the family of $\{f_i\left(\mathbf{x}|\boldsymbol{\phi}_i\right) : \boldsymbol{\phi}_i \in \mathcal{S}^d\}$ is identifiable or, equivalently, linear independent. Indeed, linear independence is the sufficient and necessary condition of identifiability (Yakowitz and Spragins 1968).

Next, we revisit the characteristics of disjoint distribution sets (Kent 1983; Browne and McNicholas 2015) to facilitate the proof in Theorem (Identifiability).

Definition (Disjoint Distribution Set). *Given distribution set $\mathcal{F}$ and $\mathcal{G}$, if no element of $\mathcal{G}$ can be formed as a linear combination of elements of $\mathcal{F}$, or vise verse, it is said that the span of $\mathcal{F}$ and $\mathcal{G}$ are disjoint, denoted as $\mathcal{F} \cap \mathcal{G} = \emptyset$.*

Additionally, if both of $\mathcal{F}$ and $\mathcal{G}$ are identifiable and $\mathcal{F} \cap \mathcal{G} = \emptyset$, $\mathcal{F}$ and $\mathcal{G}$ are said to be identifiable disjoint. The two set could belong to different families and this can be generalized to arbitrary k sets. Formally, we have:

Definition (Identifiably Mutually Disjoint). *Given a collection of k identifiable distribution sets $\{\mathcal{G}_{\gamma_i}\}_{i=1}^{k}$, the collection $\{\mathcal{G}_{\gamma_i}\}_{i=1}^{k}$ is said to be identifiably mutually disjoint if the following holds*

$$\mathcal{G}_{\gamma_i} \cap \mathcal{G}_{\gamma_j} = \emptyset \quad \forall i, j \in \{1, 2, ..., k\},\ i \neq j$$

for identifiable distribution set $\mathcal{G}_{\gamma_i}$ for all $i \in \{1, 2, ..., k\}$, where $\gamma = \{\gamma_i\}_{i=1}^{k}$ is the set of index parameters.

An important characteristic of identifiably mutually disjoint sets of $\{\mathcal{G}_{\gamma_i}\}_{i=1}^{k}$ is union identifiability, given in the *Lemma (Identifiability on Identifiably Disjoint Sets)* as follows.

Lemma (Identifiability on Identifiably Disjoint Sets). *Given an identifiably mutually disjoint set of $\{\mathcal{G}_{\gamma,\phi_\gamma}|\phi_\gamma \in \mathcal{S}_\gamma^d, \gamma \in K\}$, the union $\mathcal{F} = \bigcup_{i=1}^{k} \mathcal{G}_{\gamma,\phi_\gamma}$ is identifiable with respect to γ if there exists a total ordering $\preceq$ on K, where $\mathcal{S}_\gamma^d$ is the corresponding parameter space of dimension d, and $K = \{1, 2, ..., k\}$.*

Proof. It is straightforward to prove this lemma indeed. According to *Definition (The Identifiability of Finite Mixture)*, for any finite mixture from $\mathcal{F}$, we consider the relation as follows,

$$\sum_{i=1}^{N} \tau_i g_{\xi_i,\eta_{\gamma_i}}(x) = \sum_{i=1}^{N} \tau_i' g_{\xi_i',\eta_{\gamma_i}'}(x),$$

where g_{ξ_i} is an element of $\mathcal{F}$, i.e., the value of ξ_i is a subset of $\{\gamma_1 \preceq \cdots \preceq \gamma_k\}$. There can at most be k different identifiable set $\mathcal{G}_{\gamma,\phi_\gamma}$ with the index of $\gamma_1, \ldots, \gamma_k$. The identifiable set $\mathcal{G}_{\gamma,\phi_\gamma}$ are mutually disjoint. Hence, we have

$$\sum_{i=1}^{l} \xi_i g_{\gamma_p}(\boldsymbol{x}|\phi_{pi}) = \sum_{j=1}^{m} \tau_j g_{\gamma_q}(\boldsymbol{x}|\phi_{qj})$$

not exist for arbitrary nonnegative l, m and for arbitrary $p, q \in \{1, 2, \ldots, k\}$, otherwise it implies the linear dependence on contrary to the *Definition (Disjoint Distribution Set)* Given the total ordering $\preceq$ on K, we assume $\gamma_1 \preceq \cdots \preceq \gamma_k$ without loss of generality, and thus order the summation of Eq. () as follows:

$$\sum_{i=1}^{k} \sum_{j=1}^{m_i} \tau_{ij} g_{\gamma_i \theta_{ij}}(x) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \tau_{ij}' g_{\gamma_i',\theta_{ij}'}(x).$$

There exists a permutation $p(\cdot)$ so that ..., and thus implies the existence of a permutation for ξ with respect of γ, i.e., the union $\mathcal{F} = \bigcup_{i=1}^{k} \mathcal{G}_{\gamma,\phi_\gamma}$ is identifiable. □

C. Theorem (Identifiability)

Proof. We prove that hyperbolic communities are identifiable from user embeddings. The hyperbolic community $\boldsymbol{C}_i$ with community embedding $\boldsymbol{\mu}_i$ is given by its corresponding generalized hyperbolic distribution $p_{\mathcal{H}}(\boldsymbol{\theta}|\boldsymbol{\psi}_i)$. Additionally, with *Lemma (Positive Definite)*, finite mixture of generalized hyperbolic distribution $\sum_{i=1}^{C^x} \boldsymbol{\pi}_i p_{\mathcal{H}}(\boldsymbol{\theta}^x|\boldsymbol{\psi}_i^x)$ is never collapsed in the optimization to identify hyperbolic communities and metric of hyperbolic geometry is naturally captured in the variance matrix $\boldsymbol{\Sigma}$, a.k.a., scatter matrix. Thus, the identifiability of hyperbolic communities lies in the the identifiability of finite mixture of generalized hyperbolic distribution. According to the *Definition (Identifiability)*, it is sufficient and necessary that the family $\{p_{\mathcal{H}}(\boldsymbol{\theta}|\boldsymbol{\psi}) : \boldsymbol{\psi} \in \mathcal{S}^d\}$ in the parameter space $\mathcal{S}^d$ of dimension d is identifiable. Note that, identifiability is ensured if identifiability is shown under any one-to-one parameterization (Browne and McNicholas 2015).

Under a one-to-one parameterization of $\delta = \beta/\sigma^2, \alpha = 1/\sigma \times \sqrt{\omega + \beta^2/\sigma^2}$ and $\kappa = \sigma\sqrt{\omega}$, the density (p.d.f.) in Eq. () of univariate emerges:

$$f(x|\boldsymbol{\psi}) = \left[\frac{1 + (x-\mu)^2/\kappa^2}{1 + \delta^2/\left(\alpha^2 - \delta^2\right)}\right]^{\frac{\lambda - 1/2}{2}} \frac{\exp\{(x-\mu)\delta\}}{\sqrt{2\pi\sigma^2}} \frac{K_{\lambda - p/2}\left(\alpha\sqrt{[\kappa^2 + (x-\mu)^2]}\right)}{K_\lambda\left(\kappa\sqrt{\alpha - \delta^2}\right)}. \tag{1}$$

where we use $\boldsymbol{\psi} = (\delta, \alpha, \kappa)$ to denote the parameters. The study (Browne and McNicholas 2015) prove that the sets $\mathcal{G}_{(\alpha,\delta)}$ with the index of (α, δ) generated by the density in Eq. () are identifiable and mutually disjoint under a total ordering $\preceq$. Applying *Lemma (Identifiability on Identifiably Disjoint Sets)*, univariate of density in Eq. () is identifiable. Recall that density of Eq. () is valid in arbitrary q dimension. Accordingly, we have the multivariate of dimension p is identifiable if the identifiability of its univariate is given. It is straightforward to be proved according to Theorem 1 in the study (Holzmann, Munk, and Gneiting 2006) with the supporting point given in the study (Yakowitz and Spragins 1968). Thus, we can claim the identifiability of hyperbolic communities in HyperCom and complete the proof. □

D. Details of the Dataset

We will present the details of the dataset soon.

References

Browne, R. P., and McNicholas, P. D. 2015. A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2):176–198.

Holzmann, H.; Munk, A.; and Gneiting, T. 2006. Identifiability of finite mixtures of elliptical distributions. *Scandinavian journal of statistics* 33(4):753–763.

Kent, J. T. 1983. Identifiability of finite mixtures for directional data. *The Annals of Statistics* 984–988.

Yakowitz, S. J., and Spragins, J. D. 1968. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* 209–214.