PERFECT: a hyPERbolic embedding For joint usEr and Community alignmenT On the Theorem (Identifiability)

Anonymous Author(s)

In this appendix, we give more details on the Theorem (Identifiability). First, we introduce the preliminaries of the concept of identifiability in Section I, including its formal definition and the properties. Then, based on the preliminaries, we prove the Theorem (Identifiability) in Section II, verifying the effectiveness of the proposed optimization algorithm (Algorithm 1) in PERFECT.

I. PRELIMINARY: IDENTIFIABILITY

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

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

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

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

$$(\pi_i, \boldsymbol{\phi}_i) = \left(\pi_{p(i)}, \boldsymbol{\phi}_{p(i)}\right).$$

Evidently, $f(\boldsymbol{x}|\boldsymbol{\phi})$ is said to be identifiable if the family of $\{f_i(\boldsymbol{x}|\boldsymbol{\phi}_i) : \boldsymbol{\phi}_i \in \mathcal{S}^d\}$ is identifiable or, equivalently, linear independent. Indeed, linear independence is the sufficient and necessary condition of identifiability [1].

Next, we revisit the characteristics of disjoint distribution sets for the proof in Theorem (Identifiability).

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

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

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

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

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

An important characteristic of identifiably mutually disjoint sets of $\{\mathcal{H}_{\gamma_i}\}_{i=1}^{k}$ 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 $\{\mathcal{H}_{\gamma,\phi_\gamma} | \phi_\gamma \in \mathcal{S}^d_\gamma, \gamma \in K\}$, the union $\mathcal{F} = \bigcup_{i=1}^{k} \mathcal{H}_{\gamma,\phi_\gamma}$ is identifiable w.r.t. γ if there exists a total ordering $\preceq$ on K, where $\mathcal{S}^d_\gamma$ is the corresponding parameter space of dimension d, and number domain $K = \{1, 2, ..., k\}$.*

Proof. Utilizing the definitions above, it is straightforward to prove this lemma. We start with the simplest case of two mutually disjoint identifiable sets $\mathcal{H}_{1,\phi_1}$ and $\mathcal{H}_{2,\phi_2}$. We give the proof by contradiction. If the following equation

$$\sum_{i=1}^{l} \xi_i h_{\gamma_p}(\boldsymbol{x}|\phi_{pi}) + \sum_{j=1}^{m} \tau_j h_{\gamma_q}(\boldsymbol{x}|\phi_{qj}) = 0 \tag{1}$$

holds for arbitrary nonnegative l and m, it implies some linear dependence as both terms are nonnegative. This is a contraction because they are identifiably mutually disjoint.

According to *Definition (The Identifiability of Finite Mixture)*, for any finite mixture from $\mathcal{F}$, we consider the relation as follows,

$$\sum_{i=1}^{C} \tau_i h_{\gamma_i,\eta_{\gamma_i}}(x) = \sum_{i=1}^{C} \tau'_i h_{\gamma'_i,\eta'_{\gamma_i}}(x), \tag{2}$$

where h_{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 are at most k different identifiable set $\mathcal{H}_{\gamma,\phi_\gamma}$ with the index of $\gamma_1, \gamma_2, \ldots, \gamma_k$. Given the total ordering $\preceq$ on the domain K, we assume $\gamma_1 \preceq \cdots \preceq \gamma_k$ without loss of generality, and we reorder the summation above as follows:

$$\sum_{i=1}^{k} \sum_{j=1}^{m_i} \tau_{ij} h_{\gamma_i \theta_{ij}}(x) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \tau'_{ij} h_{\gamma'_i,\theta'_{ij}}(x). \tag{3}$$

There exists a permutation $p(\cdot)$ so that $\theta_{ij} = \theta'_{ij}$, and thus implies the existence of a permutation for γ, *i.e.*, the union of identifiably disjoint sets $\mathcal{F} = \bigcup_{i=1}^{k} \mathcal{H}_{\gamma,\phi_\gamma}$ is identifiable. $\square$

II. Proof of the Theorem

Theorem (Identifiability). *The user embeddings in the common hyperbolic subspace are identifiable w.r.t. a finite set of communities of the (generalized) hyperbolic distribution.*

Proof. We elaborate the identifiability of hyperbolic communities without the superscript as the proofs of identifiability in source network and target network are the same.

The hyperbolic community $\mathcal{C}_p$ with community embedding $\boldsymbol{\mu}_p$ is given by its corresponding generalized hyperbolic distribution $p_{\mathcal{H}}(\boldsymbol{\theta}|\boldsymbol{\psi}_p)$. With the positive definite of all scatter matrix Δ_p guaranteed, *i.e.*, *Theorem (Positive Definiteness)*, the finite mixture of generalized hyperbolic distribution $\sum_{p=1}^{C^x} \boldsymbol{\pi}_p p_{\mathcal{H}}(\boldsymbol{\theta}^x|\boldsymbol{\psi}_p^x)$ is never collapsed in the optimization of Algorithm 1. Thus, the identifiability of hyperbolic communities lies in the identifiability of finite mixture of generalized hyperbolic distribution. According to *Definition (Identifiability)*, the sufficient and necessary condition is 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.

Here, we start from the identifiability of its univariate distribution. The density of Eq. (7) in Section III-D is valid for arbitrary d dimension. 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 of univariate generalized hyperbolic distribution emerges:

$$p_{\mathcal{H}}(\theta|\boldsymbol{\psi}) = \left[\frac{1 + (\theta-\mu)^2/\kappa^2}{1 + \delta^2/\left(\alpha^2 - \delta^2\right)}\right]^{\frac{\lambda - 1/2}{2}} \frac{\exp\{(\theta-\mu)\delta\}}{\sqrt{2\pi\sigma^2}} \frac{K_{\lambda - d/2}\left(\alpha\sqrt{[\kappa^2 + (\theta-\mu)^2]}\right)}{K_\lambda\left(\kappa\sqrt{\alpha - \delta^2}\right)}. \tag{4}$$

where we use $\boldsymbol{\psi} = (\delta, \alpha, \kappa)$ to denote the parameters. The study [2] proves that the sets $\mathcal{H}_{(\alpha,\delta)}$ with the index of (α, δ) generated by the density in Eq. (4) are identifiable and mutually disjoint under a total ordering $\preceq$. Applying *Lemma (Identifiability on Identifiably Disjoint Sets)*, univariate of density in Eq. (4) is identifiable. Note that, identifiability is ensured if identifiability is shown under any one-to-one parameterization [2].

Actually, we have the multivariate hyperbolic distribution of dimension d is identifiable if the identifiability of its univariate is given. It is straightforward to be proved with the supporting point given in the study [1]. Thus, we can claim the identifiability of hyperbolic communities in PERFECT and complete the proof. □

References

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

[2] R. P. Browne and P. D. McNicholas, "A mixture of generalized hyperbolic distributions," *Canadian Journal of Statistics*, vol. 43, no. 2, pp. 176–198, 2015.