Generalized Identifiability Bounds for Mixture Models with Grouped Samples

Robert A. Vandermeulen and René Saitenmacher

***Abstract*—Recent work has shown that finite mixture models with m components are identifiable, while making no assumptions on the mixture components, so long as one has access to groups of samples of size $2m-1$ which are known to come from the same mixture component. In this work we generalize that result and show that, if every subset of k mixture components of a mixture model are linearly independent, then that mixture model is identifiable with only $(2m-1)/(k-1)$ samples per group. We further show that this value cannot be improved. We prove an analogous result for a stronger form of identifiability known as "determinedness" along with a corresponding lower bound. This independence assumption almost surely holds if mixture components are chosen randomly from a k-dimensional space. We describe some implications of our results for multinomial mixture models and topic modeling.**

***Index Terms*—Nonparametric statistics, identifiability, nonparametric mixture models, tensor factorization, topic modeling, multinomial mixture model.**

I. Introduction

FINITE mixture models have seen extensive use in statistics and machine learning. In a finite mixture model one assumes that samples are drawn according to a two-step process. First an unobserved *mixture component*, μ, is randomly selected according to a probability measure over probability measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ (δ is the Dirac measure). Next an observed sample X is drawn from μ, $X \sim \mu$. A central question in mixture modeling theory is that of *identifiability* [60]: whether $\mathcal{P}$ is uniquely determined from the distribution of X. From the law of total probability it follows that X is distributed according to $\sum_{i=1}^{m} a_i \mu_i$. Excepting trivial cases, a mixture model is not identifiable unless one makes additional assumptions about the mixture components. A standard assumption is that the mixture components $\mu_1, \ldots, \mu_m$ are elements of some parametric class of densities. A common choice for this class is the set of multivariate Gaussian distributions, which yields the well-known and frequently-used Gaussian mixture model. This model is indeed known to be identifiable [4], [14], [69]. A natural question to ask is whether it is possible for a mixture model to be identifiable without such parametric assumptions.

Robert A. Vandermeulen acknowledges support by the German Federal Ministry of Education and Research (BMBF) for the Berlin Institute for the Foundations of Learning and Data (BIFOLD) (01IS18037A). René Saitenmacher acknowledges support by the German Federal Ministry of Education and Research (BMBF) in the project Patho234 (031L0207D).

Robert A. Vandermeulen is with the Berlin Institute for the Foundations of Learning and Data and Machine Learning Group at Technische Universität Berlin.

René Saitenmacher is with the Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany. This work was done while RS was with Machine Learning Group at Technische Universität Berlin.

In [65] the authors consider an alternative setting for mixture modeling where *no* assumptions are made on the mixture components $\mu_1, \ldots, \mu_m$ and, instead of having access to a collection of samples where each sample is drawn from an unobserved mixture component $\mu \sim \mathcal{P}$, one has access to a collection of groups of n samples, where each group has the form $\mathbf{X} = (X_1, \ldots, X_n)$ with each X_i known to be independently sampled from μ, i.e. $X_1, \ldots, X_n \overset{iid}{\sim} \mu$. In [65] the authors develop several fundamental bounds relating the identifiability of $\mathcal{P}$ to the number of samples per group n and the number of mixture components m. These bounds consider extremal cases where there are either no assumptions on the mixture components or they are assumed to be linearly independent. If the mixture components are assumed to lie in a finite dimensional space, such as when the sample space is finite (a multinomial mixture model), then it is reasonable to assume that the collection of all mixture components is linearly dependent, however sufficiently small subsets of the mixture components are linearly independent.

In this paper we prove two fundamental bounds relating the identifiability of $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ to the number of mixture components m, the number of samples per group n, and a value k which describes the degree of linear independence of the mixture components. We show that if every subset of k measures in $\mu_1, \ldots, \mu_m$ are linearly independent, then $\mathcal{P}$ is the simplest mixture, in terms of the number of mixture components, yielding the distribution on $\mathbf{X}$ if $2m-1 \leq (k-1)n$. If n is even-valued and $2m-2 \leq (k-1)(n-1)$ then $\mathcal{P}$ is the *only* mixture, with any number of components, yielding the distribution on $\mathbf{X}$. We furthermore show that the first bound is tight and that the second bound is nearly tight. Most of the bounds in [65] are special cases of the bounds presented in this paper. Our bounds also generalize a classical result on the identifiability of mixture models where all mixture components are linearly independent [69]. We also show that this linear independence assumption occurs naturally, similarly to results in [37], and describe some practical implications of our results.

II. Background

We introduce the mathematical setting used in the rest of the paper before reviewing existing results.

A. Problem Setting

The setting described here is drawn from [65] and is highly general, assuming no regularity conditions on the mixture components. Let $(\Omega, \mathcal{F})$ be a measurable space, with $\mathcal{F}$ being its σ-algebra, and let $\mathcal{D}$ be the space of probability measures

on $(\Omega, \mathcal{F})$. Note that $\mathcal{D}$ is contained in the vector space of finite signed measures on $(\Omega, \mathcal{F})$, a fact which we will use often. For an element γ, let δ_γ denote the Dirac measure at γ. We equip $\mathcal{D}$ with the power σ-algebra, as we will be interested in measures of the form $\sum_{i=1}^m a_i \delta_{\mu_i}$ with $\mu_i \in \mathcal{D}$. We call a measure $\mathcal{P}$ on $\mathcal{D}$ a *mixture of measures* if it is a probability measure on $\mathcal{D}$ of the form $\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$ with $a_i > 0$, $\sum_{i=1}^m a_i = 1$, and $m < \infty$. We will always assume that the representation of a mixture of measures has minimal m, i.e. there are no repeated μ_i in the summands. For a full technical treatment of the concept of minimal representation see [65]. We refer to the measures $\mu_1, \ldots, \mu_m$ as *mixture components*. We now introduce the model we wish to investigate in this paper which is termed the *grouped sample setting* in [65]. If we let $\mu \sim \mathcal{P}$ and $X_1, \ldots, X_n \overset{iid}{\sim} \mu$ then the probability distribution for $\mathbf{X} = (X_1, \ldots, X_n)$ is $\sum_{i=1}^m a_i \mu_i^{\times n}$. With this in mind we introduce the following operator[1],

$$V_n(\mathcal{P}) \triangleq \sum_{i=1}^m a_i \mu_i^{\times n}. \tag{1}$$

To give some concreteness to this setting it can be helpful to consider the application of topic modeling with a finite number of topics. Here $\mu_1, \ldots, \mu_m$ are topics, which are simply distributions over words. The measure $\mathcal{P}$ designates a topic μ_i being chosen with probability a_i. The group of samples $(X_1, \ldots, X_n)$ represent a document containing n words as a bag of words. A collection of documents $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are then iid samples of $V_n(\mathcal{P})$ where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n})$. In this setting we are interested in the number of words necessary per document to recover the true topic model $\mathcal{P}$.

We will be investigating two forms of identifiability, *n-identifiability* where $\mathcal{P}$ is the simplest mixture of measures, in terms of the number of mixture components, yielding the distribution on $(X_1, \ldots, X_n)$, and *n-determinedness* where $\mathcal{P}$ is the *only* mixture of measures yielding the distribution on $(X_1, \ldots, X_n)$. We finish this section with the following two definitions which capture these two notions of identifiability.

Definition 2.1: A mixture of measures $\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$ is *n-identifiable* if there exists no mixture of measures $\mathcal{P}' \neq \mathcal{P}$ with m or fewer components such that $V_n(\mathcal{P}') = V_n(\mathcal{P})$.

Definition 2.2: A mixture of measures $\mathcal{P}$ is *n-determined* if there exists no mixture of measures $\mathcal{P}' \neq \mathcal{P}$ such that $V_n(\mathcal{P}') = V_n(\mathcal{P})$.

B. Previous Results

Here we recall several results from [65]. In that paper the authors prove five bounds relating identifiability or determinedness to the geometry of the mixture components, the number of mixture components m, and the number of samples per group n. For brevity we have summarized these bounds in Table I. [65] showed that none of these bounds are improvable by proving matching lower bounds. For clarity we include an example of a precise statement of an entry in this table.

[1]All power operators utilize the standard product corresponding to the space they operate in, such as the power measure or the power σ-algebra [34].

Theorem 2.1 (Table I Row Four or [65] Theorem 4.6): Let $\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$ be a mixture of measures where $\mu_1, \ldots, \mu_m$ are linearly independent. Then $\mathcal{P}$ is 4-determined.

The last row of Table I contains a property known as *joint irreducibility* which was introduced in [11]. A collection of probability measures $\mu_1, \ldots, \mu_m$ is *jointly irreducible* when all probability measures in the linear span of $\mu_1, \ldots, \mu_m$ lie in the convex hull of $\mu_1, \ldots, \mu_m$, i.e. $\operatorname{span}(\{\mu_1, \ldots, \mu_m\}) \cap \mathcal{D} = \operatorname{conv}(\{\mu_1, \ldots, \mu_m\})$. We do not use joint irreducibility anywhere else in this paper, however we note that it is a property that is stronger than linear independence.

For completeness we also include the following lemmas from [65] that demonstrate the unsurprising fact that k-identifiability and k-determinedness are, in some sense, monotonic. Each lemma encapsulates two statements, one concerning identifiability and one concerning determinedness, which we have combined for brevity.

Lemma 2.1: If a mixture of measures is n-identifiable (determined) then it is q-identifiable (determined) for all $q > n$.

Lemma 2.2: If a mixture of measures is not n-identifiable (determined) then it is not q-identifiable (determined) for any $q < n$.

Finally [65] Lemma 7.1 showed that if the sample space Ω is finite, the grouped sample setting is equivalent to a multinomial mixture model where n is the number of trials and $\mu_1, \ldots, \mu_m$ are the categorical distributions for each component. One may consider the grouped sample setting to be a generalized version of multinomial mixture models.

III. Related Work

The findings of this work reside at the convergence of several distinct subjects. In this section, we explore relevant studies and contributions from these intersecting fields. Further discussion of related works in the field of discrete data clustering is included in Section V.

A. Grouped Sample Setting

A significant amount of work regarding the grouped sample setting has focused on the setting where Ω is finite, which is equivalent to a multinomial mixture model. Some of the earliest work on identifiability was done on binomial mixture models with [60] demonstrating that binomial mixture models are identifiable if the number of trials n and the number of mixture components m satisfy $n \geq 2m - 1$. These results were extended to multinomial mixture models in [38] and [23]. [50] and [65] introduced estimators for the multinomial

TABLE I
SUMMARY OF IDENTIFIABILITY RESULTS FROM [65]. m DESIGNATES THE NUMBER OF MIXTURE COMPONENTS AND n THE NUMBER OF SAMPLES PER GROUP.

Component assumption	n bound	n-ident./det.
none	$n \geq 2m - 1$	identifiable
none	$n \geq 2m$	determined
linearly independent	$n \geq 3$	identifiable
linearly independent	$n \geq 4$	determined
jointly irreducible	$n \geq 2$	determined

components when the $n \geq 2m-1$ bound is met. Turning to the continuous setting, the paper [53] introduces a method for recovering mixture components in the grouped sample setting when the components are densities on some Euclidean space. That method is furthermore guaranteed to asymptotically recover the components whenever the mixture of measures is identifiable. In [68] the authors consider the grouped sample setting where the mixture components come from some parametric class of densities and provide results for identifiability and rates of convergence.

The grouped sample setting can be considered as a special case of a finite *exchangeable sequence* [35]. An (infinite) sequence of random variables $\xi_1, \xi_2, \ldots$ is called *exchangeable* if

$$(\xi_1, \ldots, \xi_m) \stackrel{d}{=} (\xi_{k_1}, \ldots, \xi_{k_m})$$

for every distinct subsequence $\xi_{k_1}, \ldots, \xi_{k_m}$. For an infinite sequence de Finetti's Theorem tells us that (for Borel spaces) one can always decompose the distribution of the sequence to be independent, conditioned on some other random variable in a way akin to (1), though this random variable is not necessarily discrete as in (1) [34]. This theorem does not extend to finite sequences, but there exist works investigating the grouped sample setting for continuous mixtures. One such work is [67], where the authors present rates for estimating a continuous version of $\mathcal{P}$ in the context of binomial mixture models.

B. Nonparametric Mixture Modeling

Other works have investigated nonparametric mixture models without assuming the grouped sample setting. Unlike the grouped sample setting, these methods *must* make assumptions on the mixture components to guarantee identifiability. One approach to nonparametric mixture modeling is to assume a clustering structure where mixture components are assumed to correspond to modes or concentrated regions of the probability density function (pdf) [5], [6], [18], [59], [66]. Some of these methods have identifiability guarantees so long as the true mixture components are sufficiently concentrated and separated.

Unsurprisingly, there has been significant investigation into mixture models using neural networks. The basic approach to this assumes data is generated by sampling an unobserved latent variable from a mixture model, $Z_i \overset{iid}{\sim} \sum_{j=1}^m a_j \mu_j$, and the observed data comes from passing the latent variable through a neural network $X_i = f_\theta(Z_i)$. A common approach to this uses a variational autoencoder with mixture priors [20], [25], [30], [31], [61], usually a Gaussian mixture model. In such a situation one is interested in estimating θ and the mixture parameters of the latent distribution jointly. Clearly there are some issues with identifiability when fitting f_θ and the mixture components jointly. For example, if f_θ is linear and Z is distributed according to a Gaussian mixture model, then adjusting the mixture parameters to compensate for the linear transform can effectively fit the data. This means that one could simply fix any value of θ, knowing that the same effect could be achieved by choosing corresponding mixture parameters. Remarkably it has recently been shown that, if f_θ is a leaky ReLU network and the latent distribution is a Gaussian mixture model, one has identifiability of the mixture components up to a certain class of transformations [39].

Another approach to nonparametric mixture modeling, which avoids the need for strong separation and concentration assumptions, factors a pdf into a density with rank-one components [56]. This can be expressed as

$$p(x_1, \ldots, x_d) = \sum_{i=1}^m \prod_{j=1}^d w_i p_{i,j}(x_j). \tag{2}$$

This method has its origins in analyses of discrete distributions [3], where it can be applied for clustering discrete data.

Notably, it has been observed that the structure in (2) can be used to improve nonparametric density estimation [2], [36], [56], [57], with [62]–[64] theoretically showing that this assumption eliminates the nonparametric curse of dimensionality. These methods are identifiable under specific linear independence assumptions on the collection of component marginal distributions $p_{i,j}$. This guarantee of identifiability follows from previous works concerning the uniqueness of tensor factorizations [1], [40].

C. Tensor Factorization

Identifiability with linearly independent components given $n \geq 3$ was established in [1] by way of Kruskal's (Factorization) Theorem [40]. A spectral algorithm for the estimation of models with linearly independent components can be found in [3]. An adaptation of this algorithm can be used to recover components like those in (2) [56].

A generalization of Kruskal's Theorem for d-way arrays can be found in [55], and it is related to some of the techniques we use in the proof of Theorem 4.1. However, we never employ the results from [55] in any of our proofs. Nonetheless, the theorems we prove are a natural extension of [65] using the independence setting considered in [55].

Nonnegative matrix/tensor factorization is a constrained form of tensor factorization where one enforces positivity on the components, ensuring that all elements in the factorized matrices or tensors are nonnegative [7], [54]. This method has been used for discrete mixed membership models. Identifiability is again an important consideration for nonnegative matrix factorization. A well-known property that guarantees identifiability for nonnegative matrix factorization is the "separability condition" [21], also referred to as the "anchor word" assumption [8]. This condition is the discrete analogue of joint irreducibility.

When the space Ω is finite, then $V_n(\mathcal{P})$ can be represented by a nonnegative symmetric tensor of the form $\sum_{i=1}^m a_i p_i^{\otimes n}$, where p_i are the probability vectors associated with μ_i. In this context, finding the mixture components of $\mathcal{P}$ is equivalent to finding a nonnegative symmetric factorization of $V_n(\mathcal{P})$ [16].

IV. Main Results

In this section we present the main results of this paper. They are related to a property which we call k-*independence*.

Definition 4.1: A sequence of vectors $x_1, \ldots, x_m$ is called k-*independent* if every subsequence $x_{i_1}, \ldots, x_{i_k}$ containing k vectors is linearly independent.

The concept of k-*independence* is simply a generalization of *Kruskal rank* [40] to vector spaces. We define k-independence using a *sequence* rather than a *set* of vectors so as to relate it to a matrix rank, since a matrix can have repeated columns. When $x_1, \ldots, x_m$ are distinct (as will be the case in our main theorems) we can define k-independence simply using sets and subsets. We now present the main results of this paper.

A. Identifiability and Determinedness Bounds

Below, we present two theorems that establish conditions for the identifiability of a mixture of measures. These conditions relate the k-independence of the mixture components, the number of mixture components m, and the number of samples per group n.

Theorem 4.1: Let $m \geq 2$. If $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are k-independent and it holds that $2m - 1 \leq (k-1)n$ then $\mathcal{P}$ is n-identifiable.

Theorem 4.2: Let $m \geq 2$ and n even-valued. If $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are k-independent and it holds that $2m - 2 \leq (k-1)(n-1)$ then $\mathcal{P}$ is n-determined.

In situations where the mixture components are not linearly independent, the results here can guarantee identifiability and determinedness for significantly lower values of n than were demonstrated in [65]. The applications and implications of these findings are discussed in Section V.

For Theorem 4.1 we must omit the case where $m = 1$ since this would imply that $k = 1$ which results in the inequality "$1 \leq 0$" in the theorem statement. Note that any mixture containing only one component is trivially 1-identifiable, which is accounted for by row one in Table I.

While Theorem 4.1 is very much in the same vein as the factorization result in [55], and indeed one could apply that result to recover a rough version of Theorem 4.1, proving a precise version of this theorem necessitated developing some novel tools. In particular, we needed to precisely characterize how the tensor power affects the k-independence of a collection of vectors. This characterization is developed in Lemma 6.2. Going further, in Lemma 6.3 we also developed a tight analysis of how k-independence structure changes when tensor power is applied to a collection of vectors where some subsets are k-independent and other subsets are k'-independent with $k \neq k'$.

The proof of Theorem 4.2 uses the aforementioned lemmas as well as a new induction argument. Preserving determinedness through an induction argument presents a challenge, as there is very little prior work on this property. We emphasize once more that, unlike identifiability, the determinedness property in this setting is quite new, and as such, there are limited theoretical tools available for its analysis. It is worth noting that the technique used for the base case of our proof could be applied to any even value of n, with $n = 2$ being the most obvious choice. From this starting point, one can use our induction method for $n + 2$. We remark that starting with any base case other than $n = 4$ and then applying our induction method results in a statement that is weaker than Theorem 4.2.

B. Lower Bounds

Here we present two theorems which characterize the optimality of the previous theorems. Proving both of these statements is a straightforward task, utilizing identifiability and determinedness lower bounds, along with some other tools, all of which are from [65]. For identifiability we have the following theorem.

Theorem 4.3: For all $m \geq k \geq 2$ and n with $2m - 1 > (k-1)n$ there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ where $\mu_1, \ldots, \mu_m$ are k-independent and $\mathcal{P}$ is not n-identifiable.

From this we have that Theorem 4.1 cannot be improved for any values of m, n, or k, not satisfying $2m-1 \leq (k-1)n$, and is hence completely tight.

For determinedness we have a similar bound.

Theorem 4.4: For all $m \geq k \geq 2$ and n with $2m > (k-1)n$ there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ where $\mu_1, \ldots, \mu_m$ are k-independent and $\mathcal{P}$ is not n-determined.

Unfortunately this bound does not match the result from Theorem 4.2, with the simplest setting not fitting either determinedness bound being $m = 6, n = 4, k = 4$. If $n = 2$ then Theorem 4.2 only holds with $m = 1$. Interestingly the bounds match for all valid settings, i.e. $m \geq k$, when $k = 3$. We truly have no insight into which bound is loose, and it's entirely plausible that both of them are, or that it may depend on the specific values of the parameters m, n, and k. Its not possible for 4.2 to hold for $n = 3$, due to the previous lower bound on determinedness with linearly independent components, but it is possible that it may hold for other odd-valued n.

C. Comparison to Previous Results

The results in Section IV are quite general and contain four of the five bounds from [65] as special cases. Since any pair of distinct probability measures are not collinear, it follows that the components of any mixture of measures with at least two mixture components are 2-independent. Setting $k = 2$ in Theorems 4.1 and 4.2 gives us the first two bounds from Table I (noting that n is always even for the determinedness result).

If a collection of m vectors are linearly independent we have that they are m-independent. Setting $k = m$ in Theorem 4.1 we have

$$\begin{aligned} & 2m - 1 \leq (m-1)n \\ \iff \quad & \frac{2m-1}{m-1} \leq n \\ \iff \quad & \frac{2m-2+1}{m-1} \leq n \\ \iff \quad & \frac{2m-2}{m-1} + \frac{1}{m-1} \leq n, \end{aligned}$$

with the minimal n satisfying this being 3 which yields row 3 in Table I. The analogous determinedness bound on row 4 can similarly be derived from Theorem 4.2,

$$2m - 2 \leq (m-1)(n-1) \Rightarrow 2 \leq n - 1 \iff 3 \leq n,$$

and the smallest even-valued n satisfying this bound is 4. As a final point we remark that, in contrast to previous results, Theorems 4.1 and 4.2 imply that, when $n = 3$ or $n = 4$ respectively, it is possible to have identifiability/determinedness without linearly independent components; for example by setting $n = 4$, $k = 7$, and $m = 10$ for the determinedness case.

V. Discussion

Before presenting our proofs we discuss some of the applications and implications of the main results.

A. Applications: Multinomial Mixture Modeling

The results here are perhaps most pertinent to the application of multinomial mixture models [38], which are equivalent to the grouped sample setting with mixture components on a finite sample space, $|\Omega| = d < \infty$ [65]. These models are applicable to settings where one wants to cluster subjects and has access to repeated categorical samples from each subject. This can be considered as a type of discrete data clustering problem [49]. There exist many real world applications where this model is appropriate. Using the notation from Section II-A we will describe some settings where this model has been applied.

Multinomial mixture models have been used in business analytics to cluster customer types where $\mathbf{X}_i$ are the purchases of customer i with $X_{i,j}$ representing one instance of customer i purchasing a particular product [15]. In the cognitive sciences one may be interested in finding clusters of behavior over a collection of subjects. Here, $X_{i,j}$ may represent a response to a questionnaire, an experimental observation or a physiological or psychological reading, with $\mathbf{X}_i$ denoting the collection of repeated readings for subject i over a period of time. In such scenarios, one might expect an underlying factor to be revealed through repeated measurements rather than a one-time observation. As an illustrative example, experiencing anxiety from time to time is normal, but frequently reporting anxiety might indicate an underlying condition. Similarly, only repeatedly solving a certain task might be adequate to reliably indicate a subject's understanding or ability. In this context, multinomial mixture models have been used to model the response of children in an experiment to assess their understanding of the physical world [9], [13], [24], [41]. Other instances where multinomial mixture models are used to model repeated tasks include reaction response tasks [17], [70] and experiments that study the reliance of adults on potentially misleading visual cues [28].

Elsewhere multinomial mixture modeling has been used to cluster different types of internet traffic [32]. These methods have also found use in topic modeling or text clustering [42], [46], [52], however a great deal of topic modeling focuses on mixed membership models such as Latent Dirichlet Allocation [12] or a Dirichlet-multinomial model. Multinomial mixture models are desirable when one's focus is clustering.

Dirichlet-multinomial mixtures are an alternative to multinomial mixture models that replaces the multinomial components with Dirichlet-multinomial components. Like the multinomial distribution, the Dirichlet-multinomial distribution is a distribution defined on count data. For count data with d outcomes and n trials, the Dirichlet-multinomial distribution is parameterized by a positive vector $\alpha \in \mathbb{R}^d$ and its pmf on count data $x \in \mathbb{R}^d$ is defined as

$$\text{DirMult}(x; n, \alpha) = \mathbb{E}_{p\sim\text{Dirichlet}(\alpha)}\left[\text{Mult}(x; n, p)\right], \quad (3)$$

where Mult is the multinomial distribution. A multinomial mixture model could be employed instead of a Dirichlet-multinomial mixture in applications where Dirichlet-multinomial mixtures are used, such as short text clustering [22], [43], [71] and clustering genetic/biological data [27], [29], [47]. The two models each have their own advantages and disadvantages. Specifically, a multinomial mixture model is generally easier to estimate and interpret, while the Dirichlet-multinomial mixture offers more flexibility. Dirichlet-multinomial mixtures also lack the identifiability guarantees that exist for multinomial mixture models, which can help guide data collection and estimation. In the sequel we describe how the identifiability and determiendness guarantees presented in this work can aid in statistical analysis and data collection.

B. Implications: Multinomial Mixture Modeling

In the aforementioned settings, with the samples $X_{i,j}$ taking one of d different values, it can be very natural to assume that the mixture components are d-independent due to the following proposition.

Proposition 5.1: Let Ψ be a measure which is absolutely continuous to the uniform measure on the probability simplex Δ^{d-1} and let $\Gamma_1, \ldots, \Gamma_d \overset{iid}{\sim} \Psi$. Then $\Gamma_1, \ldots, \Gamma_d$ are linearly independent with probability one.

This fact is particularly relevant for topic modeling where d, the number of words in a vocabulary, can be large and estimation can be difficult. A straightforward way to fix this is to assign words to $d' < d$ clusters, perhaps using a vector word embedding [44], thereby coarsening the event space. To recover m topics we should have $d' \leq m$ and satisfy $2m - 1 \leq (d' - 1)n$ where n is the number of words per document. To test whether a corpus could potentially contain more topics than a proposed topic model with m topics, we would need that $2m-2 \leq (d'-1)(n-1)$. When $d < m$ and the components cannot be linearly independent, the results here can guarantee identifiability with n being a $d-1$ factor smaller than was shown in [65], thus requiring less data to be collected. Alternatively the results presented here demonstrate how one may increase d, perhaps by adjusting the data collection, so that it is possible to recover the desired number of components.

Another consideration is the difficulty of estimating $V_n(\mathcal{P})$. For a given d and n, $V_n(\mathcal{P})$ has the form of a symmetric tensor in $\mathbb{R}^{d^{\times n}}$ [65]. While symmetry aids in the estimation of $V_n(\mathcal{P})$, this space of symmetric tensors has a substantial dimensionality of $\binom{n+d-1}{d}$ [16]. Thus it is desirable to choose n and d as small as possible while still meeting the needs for the analysis at hand. To reduce d, for example, one may opt to have a shorter questionnaire for psychometrics or to cluster product types for customer information, e.g., treat "apples," "pears," and "bananas" as "fruit."

Finally, the determinedness property can aid in answering the question "would we find more components if we collect more samples, n, for each subject?" Suppose $\hat{V}$ is a highly accurate estimate of $V_n(\mathcal{P})$, $\hat{\mu}_1, \ldots, \hat{\mu}_m$ is a collection of probability measures, and $\hat{a} \in \Delta^{m-1}$, such that $\hat{V} \approx \sum_{i=1}^m \hat{a}_i \hat{\mu}_i^{\times n}$. Theorem 4.2 gives us an indication that, if m is sufficiently small with respect to k and n, then all the components have been recovered. Naturally it is impossible to know to a certainty that one has found all components due to sampling error, but Theorem 4.2 can give some guidance as to whether one should consider using a larger n.

C. Other Points

The intuition from Proposition 5.1 applies to any setting where one assumes that there exists some linearly independent collection of probability measures, $\xi_1, \ldots, \xi_d$, with each component being mixture of these components, $\mu_i = \sum_{j=1}^d c_{i,j}\xi_j$. For example, one may assume that each component is mixture of a fixed but unknown set of d multivariate Gaussian distributions, with each μ_i having a different mixture weighting of the Gaussian components. Again, the advantage of our results over previous ones is the ability to choose n more conservatively.

The grouped sample setting also occurs naturally in many problem settings including group anomaly detection [45], transfer learning [10], and distribution regression/classification [48], [58]. In these settings one has access to groups of samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n})$. Mathematical analysis of techniques in this setting typically assume $n \to \infty$. The study of such problems for fixed n is less explored and the results here may help research into this setting.

VI. Proofs

This section contains proofs of the results in Section IV and supporting lemmas. Proofs omitted in this section can be found in Appendix A. The symbol $\otimes$ represents the standard tensor product on a Hilbert space and is also used in the superscript to denote tensor powers. The $\prod$ symbol represents tensor product when applied to elements of a Hilbert space[2]. For a measure μ, $\mu^{\times n}$ denotes the power measure, as induced by the standard product measure, and acts on the power σ-algebra, as induced by the standard product σ-algebra [34]. For real-valued functions f and g, $f \times g$ represents the outer product of the functions, i.e. $(f \times g)(a, b) = f(a)g(b)$; this notation is also used in the superscript to denote a power of this kind of product. For a natural number N, $[N]$ is defined to be $\{1, 2, \ldots, N\}$. To streamline the presentation of our main theorems we first introduce the mathematical tools we will be using.

The following lemma is not particularly novel, but we will be using it quite extensively without reference so we include a statement of it here.

Lemma 6.1: Let $x_1, \ldots, x_m$ nonzero be vectors in an inner product space. Then $x_1, \ldots, x_m$ are linearly independent iff there exist vectors $z_1, \ldots, z_m$ such that $\langle x_i, z_i \rangle \neq 0$ for all i and $\langle x_i, z_j \rangle = 0$ for all $i \neq j$.

[2]Some works use $\otimes_{i=1}^n$ instead of this notation.

The next lemma serves as something of a workhorse in our proofs.

Lemma 6.2: Let $x_1, \ldots, x_m$ be vectors in a Hilbert space which are k-independent with $k \geq 2$. Then $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ are $\min(n(k-1)+1, m)$-independent.

Proof: We will first consider the case where $n(k-1)+1 = m$. We can relabel the vectors $x_1, \ldots, x_{n(k-1)+1}$ as x and $x_{i,j}$ where $(i, j) \in [n] \times [k-1]$. By k-independence, for all i, there exists a vector z_i such that $\langle z_i, x \rangle = 1$ and $\langle z_i, x_{i,j} \rangle = 0$ for all j. From this we have that

$$\left\langle x^{\otimes n}, \prod_{i=1}^n z_i \right\rangle = \prod_{i=1}^n \langle x, z_i \rangle = 1, \text{ and}$$
$$\left\langle x_{i,j}^{\otimes n}, \prod_{l=1}^n z_l \right\rangle = \prod_{l=1}^n \langle x_{i,j}, z_l \rangle = 0, \text{ for all } i, j.$$

Because the relabeling is arbitrary it follows that for all $i' \in [n(k-1)+1]$ there exists $\mathbf{z}_{i'}$ such that $\left\langle x_{i'}^{\otimes n}, \mathbf{z}_{i'} \right\rangle = 1$ and $\mathbf{z}_{i'} \perp x_{j'}^{\otimes n}$ for all $j' \neq i'$. Thus we have that $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ are m-independent. We will now consider two other cases for the value of m. For $m < n(k-1)+1$ we can show that $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ are linearly independent by the same argument. If $m > n(k-1)+1$ then it follows from the $m = n(k-1)+1$ case that every subsequence of length $n(k-1)+1$ of $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ is independent, so it follows that $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ is $(n(k-1)+1)$-independent. ∎

The following lemma's proof is very similar to the proof of Lemma 6.2, but we defer it to Appendix A due to its length. Note that it recovers Lemma 6.2 by setting $k' = k$.

Lemma 6.3: Let $x_1, \ldots, x_m$ be k-independent with $k \geq 2$ and x such that $x, x_1, \ldots, x_m$ is k'-independent with $k \geq k' > 1$. Then $x^{\otimes n}, x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ is $\min(m+1, (n-1)(k-1)+k')$-independent.

To prove Theorem 4.1 we will use the following slight adaptation of Kruskal's Theorem.

Theorem 6.1 (Hilbert space extension of [40]): Let $x_1, \ldots, x_r$, $y_1, \ldots, y_r$, and $z_1, \ldots, z_r$ be elements of three Hilbert spaces $\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z$ such that $x_1, \ldots, x_r$ are r_x-independent with r_y, r_z defined similarly. Further suppose that $r_x + r_y + r_z \geq 2r + 2$. If $a_1, \ldots, a_l \in \mathcal{H}_x, b_1, \ldots, b_l \in \mathcal{H}_y$, and $c_1, \ldots, c_l \in \mathcal{H}_z$ with $r \geq l$ such that

$$\sum_{i=1}^r x_i \otimes y_i \otimes z_i = \sum_{j=1}^l a_j \otimes b_j \otimes c_j,$$

then $l = r$ and there exists a permutation $\sigma : [r] \to [r]$ and $D_x, D_y, D_z \in \mathbb{R}^r$ such that $a_{\sigma(i)} = x_i D_{x,i}, b_{\sigma(i)} = y_i D_{y,i}$, and $c_{\sigma(i)} = z_i D_{z,i}$ with $D_{x,i} D_{y,i} D_{z,i} = 1$ for all i.

The following three lemmas allow us to embed general measures in Hilbert spaces and will allow us to use tools from Hilbert space theory [33].

Lemma 6.4 (Lemma 6.2 from [65]): Let $\gamma_1, \ldots, \gamma_n$ be finite measures on a measurable space $(\Psi, \mathcal{G})$. There exists a finite measure π and nonnegative functions $f_1, \ldots, f_n \in L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ such that, for all i and all $B \in \mathcal{G}$

$$\gamma_i(B) = \int_B f_i d\pi.$$

The last lemma will be used in particular to embed collections of probability measures in a joint measure space as pdfs.

Lemma 6.5 (Lemma 6.3 from [65]): Let $(\Psi, \mathcal{G})$ be a measurable space, γ and π a pair of finite measures on that space, and f a nonnegative function in $L^1(\Psi, \mathcal{G}, \pi)$ such that, for all $A \in \mathcal{G}$, $\gamma(A) = \int_A f d\pi$. Then for all n, for all $B \in \mathcal{G}^{\times n}$ we have

$$\gamma^{\times n}(B) = \int_B f^{\times n} d\pi^{\times n}.$$

Lemma 6.6 (Lemma 5.2 from [65]): Let $(\Psi, \mathcal{G}, \gamma)$ be a measure space. There exists a unitary transform $U : L^2(\Psi, \mathcal{G}, \gamma)^{\otimes n} \to L^2(\Psi^{\times n}, \mathcal{G}^{\times n}, \gamma^{\times n})$ such that, for all $f_1, \ldots, f_n \in L^2(\Psi, \mathcal{G}, \gamma)$,

$$U(f_1 \otimes \cdots \otimes f_n) = f_1(\cdot) \cdots f_n(\cdot).$$

Finally we remind the reader of the following standard result from real analysis.

Lemma 6.7 (Proposition 2.23 from [26]): Let $(\Psi, \mathcal{G}, \gamma)$ be a measure space and $f, g \in L^1(\Psi, \mathcal{G}, \gamma)$. Then $f = g$ γ-almost everywhere iff, for all $A \in \mathcal{G}$, $\int_A f d\gamma = \int_A g d\gamma$.

For the rest of the paper we will leave the "almost everywhere" qualifier implicit. We can now prove the main theorems in Section IV.

Proof of Theorem 4.1: Let $\mathcal{Q} = \sum_{i=1}^{l} b_i \delta_{\nu_i}$ be a mixture of measures with $l \leq m$, such that $V_n(\mathcal{P}) = V_n(\mathcal{Q})$. From this we have that

$$\sum_{i=1}^{m} a_i \mu_i^{\times n} = \sum_{j=1}^{l} b_j \nu_j^{\times n}.$$

From Lemma 6.4 there exists a measure ξ and nonnegative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\xi) \cap L^2(\xi)$, such that, for all measurable A and i, $\mu_i(A) = \int_A p_i d\xi$ and $\nu_i(A) = \int_A q_i d\xi$. From Lemmas 6.5 and 6.7 we have that

$$\sum_{i=1}^{m} a_i p_i^{\times n} = \sum_{j=1}^{l} b_j q_j^{\times n},$$

and from Lemma 6.6 we have

$$\sum_{i=1}^{m} a_i p_i^{\otimes n} = \sum_{j=1}^{l} b_j q_j^{\otimes n}. \tag{4}$$

From the theorem hypothesis we know that $m \geq 2$, and trivially $k \leq m$, so we have the following

$$\begin{aligned} & 2m - 1 \leq (k-1)n \\ \Rightarrow \quad & 2m - 1 \leq (m-1)n \\ \Rightarrow \quad & \frac{2m-1}{m-1} \leq n \\ \Rightarrow \quad & 2 < n \\ \Rightarrow \quad & 3 \leq n. \end{aligned}$$

Because $n \geq 3$ it is always possible to decompose $n = n_1 + n_2 + n_3$ where n_i are all positive integers.

We will now prove the following claim which we will denote "†": if the sequences $p_1^{\otimes n_i}, \ldots, p_m^{\otimes n_i}$ (for all $i \in [3]$) are k_i-independent respectively and $k_1 + k_2 + k_3 \geq 2m + 2$ then the theorem conclusion follows. From (4) we have that

$$\sum_{i=1}^{m} a_i p_i^{\otimes n_1} \otimes p_i^{\otimes n_2} \otimes p_i^{\otimes n_3} = \sum_{j=1}^{l} b_j q_j^{\otimes n_1} \otimes q_j^{\otimes n_2} \otimes q_j^{\otimes n_3}.$$

From Theorem 6.1 we have that $l = m$, there exists $D_1, D_2, D_3 \in \mathbb{R}^m$, and a permutation $\sigma : [m] \to [m]$, such that, for all i

$$\begin{aligned} b_{\sigma(i)} q_{\sigma(i)}^{\otimes n_1} &= a_i p_i^{\otimes n_1} D_{1,i} \\ q_{\sigma(i)}^{\otimes n_2} &= p_i^{\otimes n_2} D_{2,i} \\ q_{\sigma(i)}^{\otimes n_3} &= p_i^{\otimes n_3} D_{3,i} \end{aligned} \tag{5}$$

where $D_{1,i} D_{2,i} D_{3,i} = 1$ for all i. Applying Lemma 6.6 to (5) we have that, for all i

$$\int q_{\sigma(i)}^{\times n_2} d\xi^{\times n_2} = \int D_{2,i} p_i^{\times n_2} d\xi^{\times n_2} \Rightarrow 1 = D_{2,i}.$$

So D_2 is a vector of ones and so is D_3 by the same argument. We have that D_1 is also a vector of ones since $D_{1,i} D_{2,i} D_{3,i} = 1$ for all i. Thus we have that $p_i = q_{\sigma(i)}$ for all i. Assuming that σ is the identity mapping it follows that $a_i = b_i$ and we have shown †.

Now that † has been demonstrated, to finish the proof we will show that we can decompose $n = n_1 + n_2 + n_3$ such that $p_1^{\otimes n_i}, \ldots, p_m^{\otimes n_i}$ are k_i-independent for each i with $k_1 + k_2 + k_3 \geq 2m + 2$ which will finish our proof. To continue we will split into the cases where $n \mod 3$ is 0, 1, or 2.

Case 0: We have that $n = 3n'$ for some positive integer n' and we can let $n_1 = n_2 = n_3 = n'$. Now we can reformulate our Hilbert space embedding:

$$\sum_{i=1}^{m} a_i p_i^{\otimes n} = \sum_{i=1}^{m} a_i p_i^{\otimes n'} \otimes p_i^{\otimes n'} \otimes p_i^{\otimes n'}.$$

From Lemma 6.2 we know that the tensors $p_1^{\otimes n'}, \ldots, p_m^{\otimes n'}$ are $\min((k-1)n' + 1, m)$-independent. If $\min((k-1)n' + 1, m)$ is m then it follows that $k_1 + k_2 + k_3 = 3m \geq 2m + 2$ since $m \geq 2$. If $\min((k-1)n' + 1, m) = (k-1)n' + 1$ then we have that $k_1 + k_2 + k_3 = 3(k-1)n' + 3 = (k-1)n + 3 \geq 2m + 2$ by the theorem hypothesis $(k-1)n \geq 2m - 1$.

Case 1: Here we have that $n = 3n' + 1$ and we let $n_1 = n_2 = n'$ and $n_3 = n' + 1$, so

$$\sum_{i=1}^{m} a_i p_i^{\otimes n} = \sum_{i=1}^{m} a_i p_i^{\otimes n'} \otimes p_i^{\otimes n'} \otimes p_i^{\otimes n'+1}.$$

From Lemma 6.2 we have that $k_1 = k_2 = \min(m, (k-1)n' + 1)$ and $k_3 = \min(m, (k-1)(n'+1) + 1)$. If we have that $\min(m, (k-1)n' + 1) = m$ then it follows that $\min(m, (k-1)(n'+1) + 1) = m$ and $k_1 + k_2 + k_3 = 3m \geq 2m + 2$.

If we have that $\min(m, (k-1)n' + 1) = (k-1)n' + 1$ and $\min(m, (k-1)(n'+1) + 1) = (k-1)(n'+1) + 1$ then we have that

$$\begin{aligned} k_1 + k_2 + k_3 &= 2((k-1)n' + 1) + (k-1)(n'+1) + 1 \\ &= (3n' + 1)(k-1) + 3 \\ &= n(k-1) + 3 \\ &\geq 2m + 2, \end{aligned}$$

by the theorem hypothesis.

Lastly, if we have that $\min(m,(k-1)n'+1) = (k-1)n'+1$ and $\min(m,(k-1)(n'+1)+1) = m$ with $m < (k-1)(n'+1)+1$ (for equality see the above $\min(m,(k-1)(n'+1)+1) = (k-1)(n'+1)+1$ case) then we have that

$$\begin{aligned} & m < (k-1)(n'+1)+1 \\ \Rightarrow \quad & m < (k-1)2n'+1 \\ \Rightarrow \quad & m \leq (k-1)2n' \\ \Rightarrow \quad & 2m+2 \leq 2((k-1)n'+1)+m = k_1+k_2+k_3, \end{aligned}$$

where the second inequality follows because $n' \geq 1$ and the third inequality follows because both sides of the inequality are integers.

Case 2: We have that $n = 3n'+2$, so let $n_1 = n'$ and $n_2 = n_3 = n'+1$. Now we have that

$$\sum_{i=1}^{m} a_i p_i^{\otimes n} = \sum_{i=1}^{m} a_i p_i^{\otimes n'} \otimes p_i^{\otimes n'+1} \otimes p_i^{\otimes n'+1}.$$

If $\min(m,(k-1)n'+1) = \min(m,(k-1)(n'+1)+1) = m$ and thus $m = k_1 = k_2 = k_3$, we have that $k_1+k_2+k_3 \geq 2m+2$ as in the previous cases. If we have that $k_2 = k_3 = \min(m,(k-1)(n'+1)+1) = (k-1)(n'+1)+1$ then it also follows that $k_1 = \min(m,(k-1)n'+1) = (k-1)n'+1$ and

$$k_1+k_2+k_3 = (k-1)(3n'+2)+3 = n(k-1)+3 \geq 2m+2.$$

Now if we have that $k_2 = k_3 = m$ and $k_1 = (k-1)n'+1$ then $k_1+k_2+k_3 = 2m+(k-1)n'+1$ and since $k \geq 2$ and $n' \geq 1$ we have that $k_1+k_2+k_3 \geq 2m+2$ so we are done, and with this final case have finished the proof. ∎

Proof Sketch of Theorems 4.3 and 4.4: Theorem 4.2 in [65] states that, for all $m \geq 2$, there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ which is not $2m-2$-identifiable. Therefore there exists a mixture of measures $\mathcal{Q} = \sum_{i=1}^{l} b_i \delta_{\nu_i}$ with $\mathcal{P} \neq \mathcal{Q}$, and $l \leq m$ such that

$$\sum_{i=1}^{m} a_i \mu_i^{\times 2m-2} = \sum_{j=1}^{l} b_j \nu_j^{\times 2m-2}.$$

Since $(k-1)n \leq 2m-2$ we have that either $(k-1)n = 2m-2$ or, if $(k-1)n < 2m-2$ we can apply Lemma 2.2, giving us

$$\sum_{i=1}^{m} a_i \mu_i^{\times (k-1)n} = \sum_{j=1}^{l} b_j \nu_j^{\times (k-1)n}.$$

Since any pair of distinct probability measures are linearly independent, it follows that any collection of probability measures are 2-independent. Using this fact we can can adapt Lemma 6.2 to show that $\mu_1^{\times k-1}, \ldots, \mu_m^{\times k-1}$ are k-independent. Letting $\mathcal{P}' = \sum_{i=1}^{m} a_i \delta_{\mu_i^{\times k-1}}$ and $\mathcal{Q}' = \sum_{i=1}^{l} b_i \delta_{\nu_i^{\times k-1}}$, we have that $V_n(\mathcal{P}') = V_n(\mathcal{Q}')$ and we are done. The proof of Theorem 4.4 is virtually identical and follows from [65] Theorem 4.4. ∎

Proof of Theorem 4.2: If $n = 2$ then

$$\begin{aligned} & 2m-2 \leq (k-1)(n-1) \\ \Rightarrow \quad & 2m-1 \leq k \leq m \quad \text{(noting } k \leq m\text{)} \\ \Rightarrow \quad & m \leq 1 \\ \Rightarrow \quad & m = 1. \end{aligned}$$

Since the theorem hypothesis assumes $m \geq 2$ we have that the theorem is vacuously true for the $n = 2$ case[3]. To finish the proof we will show that theorem holds for $n = 4$ and then proceed by induction.

Base Step: Let $n = 4$. We will proceed by contradiction and assume there exists $k \geq 2$ and a collection of k-independent probability measures $\mu_1, \ldots, \mu_m$ such that there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ and $\mathcal{Q} = \sum_{i=1}^{l} b_i \delta_{\nu_i}$ a mixture of measures where $\mathcal{P} \neq \mathcal{Q}$ and $V_4(\mathcal{P}) = V_4(\mathcal{Q})$. From the theorem hypothesis that $m \geq 2$ it follows that $k \geq 2$ and $k-2 \geq 0$. Applying this bound to our theorem hypothesis with $n = 4$ we have that

$$\begin{aligned} & 2m-2 \leq (k-1)3 \\ \Rightarrow \quad & 2m-1 \leq 3k-2 \\ & \leq 3k-2+(k-2) \\ & = 4k-4 \\ & = (k-1)4 \\ & = (k-1)n. \end{aligned}$$

Since we have that $2m-1 \leq (k-1)n$ we can apply Theorem 4.1 and thus $l > m$. We will proceed analogously to the proof of Theorem 4.1 and embed the measures in a Hilbert space as before

$$\sum_{i=1}^{m} a_i p_i^{\otimes 4} = \sum_{j=1}^{l} b_j q_j^{\otimes 4}. \tag{6}$$

Because $l > m$ there exists i such that $q_i \neq p_j$ for all j. We will assume without loss of generality that q_1 satisfies this. Let k' be the largest value such that $q_1, p_1, \ldots, p_m$ are k'-independent. We will now show that that $k' < k$. To see this suppose that $k' \geq k$ which would imply that $q_1, p_1, \ldots, p_m$ are k-independent. Observe that $m > 2$ (thus $m \geq 3$, we use this at (7)). Were this not the case then the components of $\mathcal{P}$, μ_1 and μ_2, would be linearly independent and $\mathcal{P}$ would be 4-determined from Table I row 4, thereby violating the contradiction hypothesis. With the base case $n = 4$ in our theorem hypothesis, and the fact that k and m are positive integers, we get

$$\begin{aligned} 2m-2 \leq 3(k-1) &\Rightarrow 2m \leq 3k-1 \\ &\Rightarrow \frac{4}{3}m \leq 2k-\frac{2}{3} \\ &\Rightarrow m \leq 2k-\frac{2}{3}-\frac{1}{3}m \\ &\Rightarrow m \leq 2k-\frac{2}{3}-1 \quad (m \geq 3) \qquad (7) \\ &\Rightarrow m \leq 2k-2 \quad (m \text{ is an integer}) \\ &\Rightarrow m+1 \leq 2(k-1)+1. \qquad (8) \end{aligned}$$

From application of Lemma 6.2 it follows that $q_1^{\otimes 2}, p_1^{\otimes 2}, \ldots, p_m^{\otimes 2}$ are $\min(2(k-1)+1, m+1)$-independent and from (8) it follows that they are linearly independent.

[3]It's worth noting that one does indeed have determinedness for $n = 2$ and $m = 1$ from Table I row 2.

Now we have that there exists z such that $z \perp p_i^{\otimes 2}$ for all i but $\left\langle z, q_1^{\otimes 2} \right\rangle = 1$ and thus

$$0 = \left\langle z^{\otimes 2}, \sum_{i=1}^{m} a_i p_i^{\otimes 4} \right\rangle = \left\langle z^{\otimes 2}, \sum_{j=1}^{l} b_j q_j^{\otimes 4} \right\rangle = \sum_{j=1}^{l} b_j \left\langle z, q_j^{\otimes 2} \right\rangle^2 > 0, \tag{9}$$

a contradiction. So $k' < k$.

Because $k' < k$ there must exist a collection of k' elements of $p_1, \ldots, p_m$, which we denote $p_{i_1}, \ldots, p_{i_{k'}}$, such that $q_1, p_{i_1}, \ldots, p_{i_{k'}}$ are linearly dependent; for convenience we will assume without loss of generality that these elements are $p_1, \ldots, p_{k'}$. Because $p_1, \ldots, p_{k'}$ are linearly independent but $q_1, p_1, \ldots, p_{k'}$ are linearly dependent we have that $q_1 = \sum_{i=1}^{k'} \alpha_i p_i$ for some $\alpha_1, \ldots, \alpha_{k'}$.

We will now show that $k' \geq 2k - m + 1$. Suppose this were not the case and $k' \leq 2k - m$ or equivalently $m \leq 2k - k'$. By k-independence there exists a vector z such that $\langle z, p_{k'} \rangle = 1$ with $z \perp p_1, \ldots, p_{k'-1}, p_{k'+1}, \ldots, p_k$ and another vector z' such that $\langle z', p_1 \rangle = 1$ and $z' \perp p_2, \ldots, p_{k'}, p_{k+1}, \ldots, p_m$. We know z' exists since $k' \leq 2k - m$ so the cardinality of $p_2, \ldots, p_{k'}, p_{k+1}, \ldots, p_m$ satisfies the following

$$\begin{aligned} |\{2, \ldots, k'\}| + |\{k+1, \ldots, m\}| &= k' - 1 + m - k \\ &\leq 2k - m - 1 + m - k \\ &= k - 1. \end{aligned}$$

Now we have that

$$\begin{aligned} &\left\langle z^{\otimes 2} \otimes z'^{\otimes 2}, \sum_{i=1}^{m} a_i p_i^{\otimes 4} \right\rangle \\ &= \sum_{i=1}^{m} a_i \left\langle z, p_i \right\rangle^2 \left\langle z', p_i \right\rangle^2 \\ &= \sum_{i \in \{1, \ldots, k'-1, k'+1, \ldots, k\}} a_i \left\langle z, p_i \right\rangle^2 \left\langle z', p_i \right\rangle^2 \\ &\quad + \sum_{i \in \{k', k+1, \ldots, m\}} a_i \left\langle z, p_i \right\rangle^2 \left\langle z', p_i \right\rangle^2 \\ &= \sum_{i \in \{1, \ldots, k'-1, k'+1, \ldots, k\}} a_i 0 \left\langle z', p_i \right\rangle^2 \\ &\quad + \sum_{i \in \{k', k+1, \ldots, m\}} a_i \left\langle z, p_i \right\rangle^2 0 \\ &= 0. \end{aligned}$$

On the other hand we have that

$$\begin{aligned} &\left\langle z^{\otimes 2} \otimes z'^{\otimes 2}, \sum_{i=1}^{l} b_i q_i^{\otimes 4} \right\rangle \\ &= \sum_{i=1}^{l} b_i \left\langle z, q_i \right\rangle^2 \left\langle z', q_i \right\rangle^2 \\ &\geq b_1 \left\langle z, q_1 \right\rangle^2 \left\langle z', q_1 \right\rangle^2 \\ &= b_1 \left\langle z, \sum_{i=1}^{k'} \alpha_i p_i \right\rangle^2 \left\langle z', \sum_{i=1}^{k'} \alpha_i p_i \right\rangle^2 \\ &= b_1 \left(\sum_{i=1}^{k'} \alpha_i \left\langle z, p_i \right\rangle \right)^2 \left(\sum_{i=1}^{k'} \alpha_i \left\langle z', p_i \right\rangle \right)^2 \\ &= b_1 \alpha_{k'}^2 \alpha_1^2 > 0 \end{aligned}$$

which contradicts (6). Thus we have that $k' \geq 2k - m + 1$.

We are now going to show that $q_1^{\otimes 2}, p_1^{\otimes 2}, \ldots, p_m^{\otimes 2}$ are linearly independent via Lemma 6.3. To do this we will show that $(2-1)(k-1) + k' \geq m + 1$. Using $k' \geq 2k - m + 1$ we have that

$$\begin{aligned} &2m - 2 \leq (4-1)(k-1) \quad \text{(base case hypothesis, } n = 4\text{)} \\ \Rightarrow \quad &2m \leq 3k - 1 \\ \Rightarrow \quad &m + 1 \leq 3k - m \\ \Rightarrow \quad &m + 1 \leq (k-1) + (2k - m + 1) \\ \Rightarrow \quad &m + 1 \leq (k-1) + k'. \end{aligned}$$

Since $q_1^{\otimes 2}, p_1^{\otimes 2}, \ldots, p_m^{\otimes 2}$ are linearly independent we can finish our contradiction using the same argument as in (9).

Induction Step: We will now proceed by induction along n in an increment of 2 since the theorem statement holds for even-valued n. For our inductive hypothesis assume that for even valued $n \geq 4$ and all k, m with $2m - 2 \leq (n-1)(k-1)$ that any mixture of measures with m components which are k-independent are n-determined. Consider some mixture of measures $\mathcal{P} = \sum_{i=1}^{m'} a_i \delta_{\mu_i}$ with k-independent components and $2m' - 2 \leq (k-1)((n+2)-1)$. If $k = m'$ then we have that the components are linearly independent so it is $(n+2)$-determined by Lemma 2.1 and Table I row four. Now suppose that $m' > k$. Let $\mathcal{Q}$ be a mixture of measures with l components such that $V_{n+2}(\mathcal{P}) = V_{n+2}(\mathcal{Q})$. Embedding $V_{n+2}(\mathcal{P})$ and $V_{n+2}(\mathcal{Q})$ as before we have that

$$\sum_{i=1}^{m'} a_i p_i^{\otimes n+2} = \sum_{j=1}^{l} b_j q_j^{\otimes n+2}.$$

By k-independence there exists z such that $\langle z, p_1 \rangle \neq 0$ and

$z \perp p_{m'-(k-1)+1}, \ldots, p_{m'}$. Now we have that

$$\begin{aligned}
&\sum_{i=1}^{m'} a_i p_i^{\otimes n+2} = \sum_{j=1}^{l} b_j q_j^{\otimes n+2} \\
\Rightarrow \quad &\sum_{i=1}^{m'} a_i p_i^{\otimes n} \left\langle p_i^{\otimes 2}, \cdot \right\rangle = \sum_{j=1}^{l} b_j q_j^{\otimes n} \left\langle q_j^{\otimes 2}, \cdot \right\rangle \qquad (10) \\
\Rightarrow \quad &\sum_{i=1}^{m'} a_i p_i^{\otimes n} \left\langle p_i^{\otimes 2}, z^{\otimes 2} \right\rangle = \sum_{j=1}^{l} b_j q_j^{\otimes n} \left\langle q_j^{\otimes 2}, z^{\otimes 2} \right\rangle \\
\Rightarrow \quad &\sum_{i=1}^{m'-(k-1)} a_i p_i^{\otimes n} \left\langle p_i, z \right\rangle^2 = \sum_{j=1}^{l} b_j q_j^{\otimes n} \left\langle q_j, z \right\rangle^2 .
\end{aligned}$$

The implication (10) follows from the equivalence between tensor products and Hilbert-Schmidt operators (see [33] Proposition 2.6.9). Let $\lambda = \sum_{i=1}^{m'-(k-1)} a_i \left\langle p_i, z \right\rangle^2$, $a_i' = a_i \left\langle p_i, z \right\rangle^2 / \lambda$, and $b_i' = b_i \left\langle q_i, z \right\rangle^2 / \lambda$. Without loss of generality we will assume that $\langle q_i, z \rangle \neq 0$ for $i \in [l']$ and $\langle q_i, z \rangle = 0$ for $i > l'$, with l' potentially equaling l. Note that $a_i' \geq 0$ and $\sum_{i=1}^{m'-(k-1)} a_i' = 1$ and likewise for b_i', since the right hand side of (11) is a convex combination of pdfs that itself must be equal to a pdf. So now we have

$$\sum_{i=1}^{m'-(k-1)} a_i' p_i^{\otimes n} = \sum_{j=1}^{l'} b_j' q_j^{\otimes n}. \qquad (11)$$

Note that

$$\begin{aligned}
&2m' - 2 \leq (k-1)((n+2)-1) \\
\Rightarrow \quad &2(m'-1) \leq 2(k-1) + (k-1)(n-1) \\
\Rightarrow \quad &2(m'-(k-1)) - 2 \leq (k-1)(n-1)
\end{aligned}$$

so by the induction hypothesis we have that the mixture of measures $\sum_{i=1}^{m'-(k-1)} a_i' \delta_{\mu_i}$ is n-determined[4]. It follows that $\sum_{i=1}^{m'-(k-1)} a_i' \delta_{\mu_i} = \sum_{j=1}^{l'} b_j' \delta_{\nu_j}$. Without loss of generality we will assume that $\mu_1 = \nu_1$ and $a_1' = b_1'$. It follows that $p_1 = q_1$ and thus $a_1 = b_1$. By the same argument it follows that $\nu_i = \mu_i$ and $a_i = b_i$ for all i and, because $\sum_{i=1}^{m'} b_i = 1$, that $m' = l$ and thus $\mathcal{P} = \mathcal{Q}$ and $\mathcal{P}$ is $n+2$-determined, which finishes our proof. ∎

VII. Conclusion

In this paper we have generalized previous bounds on identifiability for nonparametric mixture models with grouped samples. Per Proposition 5.1 these bounds likely better capture the typical behavior of most grouped sample settings, especially when the sample space is finite. These bounds also offer a useful guideline on how to reduce a sample space while still preserving mixture components, for example in topic modeling.

[4]There is a somewhat suble point here that two measures are equal if they admit the same measure. So even though some components in the mixture of measures $\sum_{i=1}^{m'-(k-1)} a_i' \delta_{\mu_i}$ may have a zero coefficient, it is still a mixture of measures in the sense that was described in Section II-A, although it may have fewer than $m' - (k-1)$ mixture components.

Appendix
Proofs

Proof of Lemma 6.1: For the forward direction, since $x_1, \ldots, x_m$ are linearly independent we can find the associated $z_1, \ldots, z_m$ from the Gram-Schmidt process. We prove the other direction by contradiction: suppose $x_1, \ldots, x_m$ are not linearly independent but there exist $z_1, \ldots, z_m$ satisfying the property in the lemma statement. From this it follows (without loss of generality) that $x_1 = \sum_{i=2}^{m} \alpha_i x_i$. We also know that there exists z_1 such that $\langle x_1, z_1 \rangle = 1$ but $\langle x_i, z_1 \rangle = 0$ for all $i \geq 2$. Then we have that

$$1 = \langle x_1, z_1 \rangle = \left\langle \sum_{i=2}^{m} \alpha_i x_i, z_1 \right\rangle = \sum_{i=2}^{m} \alpha_i \langle x_i, z_1 \rangle = 0,$$

a contradiction. ∎

Proof of Lemma 6.3: We first prove the case where $m+1 = (n-1)(k-1)+k'$ or equivalently $m = (n-1)(k-1)+(k'-1)$. We will use Lemma 6.1 to demonstrate the linear independence of $x, x_1, \ldots, x_m$. First we will show that there exists a tensor which is perpendicular to $x^{\otimes n}$ and all but one of the vectors in $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$. To do this we relabel $x_1, \ldots, x_m$ to $x_{i,j}$ for $(i,j) \in [n-1] \times [k-1]$ and $x_1', \ldots, x_{k'-1}'$. From k-independence we can find $z_1, \ldots, z_{n-1}$ such that $\langle z_i, x_{i,j} \rangle = 0$ for all i, j and $\langle z_i, x_1' \rangle = 1$. Likewise, from k'-independence there exists z such that $\langle z, x_1' \rangle = 1$, $\langle z, x_i' \rangle = 0$ for all $2 \leq i \leq k'-1$, and $\langle z, x \rangle = 0$. Now we have that

$$\begin{aligned}
\left\langle {x'}_1^{\otimes n}, z \otimes \prod_{i=1}^{n-1} z_i \right\rangle &= \langle x_1', z \rangle \prod_{i=1}^{n-1} \langle x_1', z_i \rangle = 1 \\
\left\langle {x'}_i^{\otimes n}, z \otimes \prod_{j=1}^{n-1} z_j \right\rangle &= \langle x_i', z \rangle \prod_{j=1}^{n-1} \langle x_i', z_j \rangle = 0 \quad \forall i \geq 2 \\
\left\langle x_{i,j}^{\otimes n}, z \otimes \prod_{l=1}^{n-1} z_l \right\rangle &= \langle x_{i,j}, z \rangle \prod_{i=1}^{n-1} \langle x_{i,j}, z_l \rangle = 0 \quad \forall (i,j) \\
\left\langle x^{\otimes n}, z \otimes \prod_{i=1}^{n-1} z_i \right\rangle &= \langle x, z \rangle \prod_{i=1}^{n-1} \langle x, z_i \rangle = 0.
\end{aligned}$$

Because x_1' was arbitrary due to relabeling, there exist tensors $\mathbf{z}_1, \ldots, \mathbf{z}_m$ such that $\left\langle x_i^{\otimes n}, \mathbf{z}_i \right\rangle = 1$ for all i, $\left\langle x_j^{\otimes n}, \mathbf{z}_i \right\rangle = 0$ for all $j \neq i$, and $\left\langle x^{\otimes n}, \mathbf{z}_i \right\rangle = 0$ for all i.

We will now find a tensor which is perpendicular to all $x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ but is not perpendicular to $x^{\otimes n}$, which will complete our proof of the $m+1 = (n-1)(k-1)+k'$ case. If $x^{\otimes n-1} \notin \operatorname{span}\left(\left\{x_1^{\otimes n-1}, \ldots, x_{(n-1)(k-1)+1}^{\otimes n-1}\right\}\right)$ then there exists a vector $\mathbf{z}$ such that $\left\langle x^{\otimes n-1}, \mathbf{z} \right\rangle = 1$ and $\left\langle x_i^{\otimes n-1}, \mathbf{z} \right\rangle = 0$ for all $i \in [(n-1)(k-1)+1]$. By k'-independence there exists a vector z such that $\langle x, z \rangle = 1$ and $\langle x_j, z \rangle = 0$ for all $j \in \{(n-1)(k-1)+1, \ldots, (n-1)(k-1)+k'-1\}$. From this it follows that $\langle x^{\otimes n}, \mathbf{z} \otimes z \rangle = 1$ but $\left\langle x_i^{\otimes n}, \mathbf{z} \otimes z \right\rangle = 0$ for all i.

Now we will assume that $x^{\otimes n-1}$ is an element of $\operatorname{span}\left(\left\{x_1^{\otimes n-1}, \ldots, x_{(n-1)(k-1)+1}^{\otimes n-1}\right\}\right)$. Note that $x_1^{\otimes n-1}, \ldots, x_{(n-1)(k-1)+1}^{\otimes n-1}$ are linearly independent by Lemma 6.2. From this it follows that there exists exactly one linear combination of $x_1^{\otimes n-1}, \ldots, x_{(n-1)(k-1)+1}^{\otimes n-1}$ which

is equal to $x^{\otimes n-1}$. We assume without loss of generality that $x^{\otimes n-1}_{(n-1)(k-1)+1}$ has a nonzero coefficient in that solution. Now have that $x^{\otimes n-1}, x_1^{\otimes n-1}, \ldots, x^{\otimes n-1}_{(n-1)(k-1)}$ are linearly independent. From this there exists $\mathbf{z}$ such that $\left\langle \mathbf{z}, x_i^{\otimes n-1} \right\rangle = 0$ for all $i \in [(n-1)(k-1)]$ but $\left\langle \mathbf{z}, x^{\otimes n-1} \right\rangle = 1$. By k'-independence there exists z such that $z \perp x_{(n-1)(k-1)+1}, \ldots, x_{(n-1)(k-1)+(k'-1)}$ but $\langle z, x \rangle = 1$. We now have that $z \otimes \mathbf{z} \perp x_i^{\otimes n}$ for all i and $\langle z \otimes \mathbf{z}, x^{\otimes n} \rangle = 1$ which finishes the $m+1 = (n-1)(k-1)+k'$ case.

We will now take care of the other cases for the values of m. If $m+1 < (n-1)(k-1)+k'$ then $x^{\otimes n}, x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ are linearly independent by the same argument made in the $m+1 = (n-1)(k-1)+k'$ case so $x^{\otimes n}, x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ are $m+1$-independent.

If $m+1 > (n-1)(k-1)+k'$ we would like to show that any subsequence of $x^{\otimes n}, x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ containing $(n-1)(k-1)+k'$ vectors is linearly independent. We consider two cases, where a subsequence contains $x^{\otimes n}$ or it does not. If it does then $x^{\otimes n}, x_{i_1}^{\otimes n}, \ldots, x_{i_{(n-1)(k-1)+k'-1}}^{\otimes n}$ is linearly independent from the $m+1 = (n-1)(k-1)+k'$ case and replacing $x^{\otimes n}$ with $x_{i_{(n-1)(k-1)+k'}}^{\otimes n}$ leaves this sequence linearly independent since $k \geq k'$ so $x^{\otimes n}, x_1^{\otimes n}, \ldots, x_m^{\otimes n}$ is $((n-1)(k-1)+k')$-independent, thus finishing the proof. ∎

Proof of Theorem 6.1: Note that two Hilbert spaces with the same finite dimension are isometric to one another. Suppose that $a_1, \ldots, a_l$, $b_1, \ldots, b_l$ and $c_1, \ldots, c_l$ with $m \leq l$ such that

$$\sum_{i=1}^{r} x_i \otimes y_i \otimes z_i = \sum_{j=1}^{l} a_j \otimes b_j \otimes c_j.$$

Let $\overline{\mathcal{H}}_x = \operatorname{span}(\{x_1, \ldots, x_r, a_1, \ldots a_l\})$ with $\overline{\mathcal{H}}_y$ and $\overline{\mathcal{H}}_z$ defined similarly. Because these spaces are finite dimensional the theorem follows from direct application of Kruskal's Theorem, see the following. ∎

Definition A.1: A matrix M has *Kruskal rank* k if every collection of k columns of M are linearly independent.

The following theorem is a statement Kruskal's Theorem [40] adapted from [51].

Theorem A.1 (Kruskal's Theorem): For a matrix M let M_i be its ith column vector. Let A, B, C be matrices of dimensions $d_A \times r, d_B \times r$, and $d_C \times r$ respectively with Kruskal rank k_A, k_B, k_C respectively and let $k_A + k_B + k_C \geq 2r+2$. Let F, G, H be matrices with dimensions $d_A \times s, d_B \times s, d_C \times s$ with $s \leq r$ and

$$\sum_{i=1}^{r} A_i \otimes B_i \otimes C_i = \sum_{j=1}^{s} F_j \otimes G_j \otimes H_j.$$

Then there exists a permutation matrix P and invertible diagonal matrices D_A, D_B, D_C such that $D_A D_B D_C = I_r$ such that

$$F = AD_AP$$
$$G = BD_BP$$
$$H = CD_CP.$$

Proof of Proposition 5.1: Let $\Gamma_1, \ldots, \Gamma_d \overset{iid}{\sim} \Psi$. We will proceed by contradiction and assume that $\Gamma_1, \ldots, \Gamma_d$ are linearly dependent with nonzero probability. It follows that that $\Gamma_1 = \sum_{i=2}^{d} \alpha_i \Gamma_i$ for some $\alpha_2, \ldots, \alpha_d$ with nonzero probability. Let 1_d be the d-dimensional vector containing all ones. Because $\Gamma_1, \ldots, \Gamma_d$ are all probability vectors it follows that

$$1_d^T \Gamma_1 = 1_d^T \sum_{i=2}^{d} \alpha_i \Gamma_i \Rightarrow 1 = \sum_{i=2}^{d} \alpha_i.$$

The probabilistic simplex lies in an affine subspace of dimension $d-1$ which we will call S. Because of this there exists an affine operator f which is a bijection between S and a closed subset of $\mathbb{R}^{d-1}$ with $f(x) = Mx + b$ for some matrix M and vector b. Let $\widetilde{\Gamma}_i = f(\Gamma_i)$. We have that $\widetilde{\Gamma}_i \sim \Psi\left(f^{-1}(\cdot)\right)$ is a measure on $\mathbb{R}^{d-1}$ which is absolutely continuous wrt the Lebesgue measure on $\mathbb{R}^{d-1}$ and thus $\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_d$ lie in general position with probability one (see [19] Section 4.5). Note that

$$\begin{aligned}\widetilde{\Gamma}_1 = M\Gamma_1 + b &= M\left(\sum_{i=2}^{d} \alpha_i \Gamma_i\right) + \left(\sum_{j=2}^{d} \alpha_j\right) b \\ &= \sum_{i=2}^{d} \alpha_i M \Gamma_i + \alpha_i b \\ &= \sum_{i=2}^{d} \alpha_i \left(M\Gamma_i + b\right) = \sum_{i=2}^{d} \alpha_i \widetilde{\Gamma}_i.\end{aligned}$$

Since $\widetilde{\Gamma}_2, \ldots, \widetilde{\Gamma}_d$ trivially lie in a $(d-2)$-dimensional affine subspace there exists a vector $v \neq 0_{d-1}$ and r such that $v^T \widetilde{\Gamma}_i = r$ for $i \geq 2$. Now we have that

$$v^T \sum_{i=2}^{d} \alpha_i \widetilde{\Gamma}_i = \sum_{j=1}^{d} \alpha_i r \Rightarrow v^T \widetilde{\Gamma}_1 = r$$

and thus, with nonzero probability, $\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_d$ do not lie in general position, a contradiction. ∎

Proof of Theorem 4.3: From [65] Theorem 4.2, for all $m \geq 2$ there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ which is not $2m-2$-identifiable, thus there exists a mixture of measures $\mathcal{Q} = \sum_{i=1}^{m'} b_i \delta_{\nu_i} \neq \mathcal{P}$ with $m' \leq m$ and

$$\sum_{i=1}^{m} a_i \mu_i^{\times 2m-2} = \sum_{j=1}^{m'} b_j \nu_j^{\times 2m-2}.$$

Since $(k-1)n \leq 2m-2$ we have that either $(k-1)n = 2m-2$ and it directly follows that

$$\sum_{i=1}^{m} a_i \mu_i^{\times (k-1)n} = \sum_{j=1}^{m'} b_j \nu_j^{\times (k-1)n}$$

or that $(k-1)n < 2m-2$ and we have

$$\begin{aligned}
&\sum_{i=1}^{m} a_i \mu_i^{\times 2m-2} = \sum_{j=1}^{m'} b_j \nu_j^{\times 2m-2} \\
\Rightarrow &\sum_{i=1}^{m} a_i \mu_i^{\times (k-1)n} \times \mu_i^{\times 2m-2-(k-1)n} \\
&\quad = \sum_{j=1}^{m'} b_j \nu_j^{\times (k-1)n} \times \nu_j^{\times 2m-2-(k-1)n} \\
\Rightarrow &\sum_{i=1}^{m} a_i \mu_i^{\times (k-1)n} \times \mu_i^{\times 2m-2-(k-1)n} \left(\Omega^{\times 2m-2-(k-1)n}\right) \\
&\quad = \sum_{j=1}^{m'} b_j \nu_j^{\times (k-1)n} \times \nu_j^{\times 2m-2-(k-1)n} \left(\Omega^{\times 2m-2-(k-1)n}\right) \\
\Rightarrow &\sum_{i=1}^{m} a_i \mu_i^{\times (k-1)n} = \sum_{j=1}^{m'} b_j \nu_j^{\times (k-1)n}.
\end{aligned}$$

If we let $\mathcal{P}' = \sum_{i=1}^{m} a_i \delta_{\mu_i^{\times k-1}}$ and $\mathcal{Q}' = \sum_{i=1}^{m'} b_i \delta_{\nu_i^{\times k-1}}$ then we have that $V_n(\mathcal{P}') = V_n(\mathcal{Q}')$. To finish the proof we will show that $\mu_1^{\times k-1}, \ldots, \mu_m^{\times k-1}$ are k-independent and we are done. To do this we will proceed by contradiction, suppose that they are not k-independent and there exists a nontrivial linear combination of k elements in $\mu_1, \ldots, \mu_m$ which is equal to zero. We will assume $\mu_1, \ldots, \mu_k$ without loss of generality satisfy this, so

$$\sum_{i=1}^{k} \alpha_i \mu_i^{\times k-1} = 0$$

and there exists i such that $\alpha_i \neq 0$. Embedding these measures as was done in the proof of Theorem 4.1 we have that

$$\sum_{i=1}^{k} \alpha_i p_i^{\otimes k-1} = 0$$

but since any pair of distinct p_i, p_j are 2-independent, applying Lemma 6.2 gives us that $p_1^{\otimes k-1}, \ldots, p_k^{\otimes k-1}$ are linearly independent, a contradiction. ∎

Proof of Theorem 4.4: From [65] Theorem 4.4, for all $m \geq 1$ there exists a mixture of measures $\mathcal{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ which is not $2m-1$-determined. From here this proof proceeds exactly as the proof of Theorem 4.3. ∎

REFERENCES

[1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.

[2] Magda Amiridi, Nikos Kargas, and Nicholas D. Sidiropoulos. Low-rank characteristic tensor density estimation part i: Foundations. *IEEE Transactions on Signal Processing*, 70:2654–2668, 2022.

[3] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[4] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: The blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of The 27th Conference on Learning Theory*, pages 1135–1164, 2014.

[5] Bryon Aragam, Chen Dan, Eric P. Xing, and Pradeep Ravikumar. Identifiability of nonparametric mixture models and Bayes optimal clustering. *The Annals of Statistics*, 48(4):2277 – 2302, 2020.

[6] Bryon Aragam and Ruiyi Yang. Uniform Consistency in Nonparametric Mixture Models. *arXiv e-prints*, page arXiv:2108.14003, August 2021.

[7] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 145–162, New York, NY, USA, 2012. ACM.

[8] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models – going beyond svd. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 1–10, Washington, DC, USA, 2012. IEEE Computer Society.

[9] Tatiana Benaglia, Didier Chauveau, and David R Hunter. An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[10] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 2178–2186, 2011.

[11] Gilles Blanchard and Clayton Scott. Decontamination of mutually contaminated models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 1–9, 2014.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[13] Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Nonparametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1):211–229, 2016.

[14] C. Bruni and G. Koch. Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.*, 13(4):1341–1357, 11 1985.

[15] Igor V Cadez, Padhraic Smyth, Edward Ip, and Heikki Mannila. Predictive profiles for transaction data using finite mixture models. Technical Report 01–67, Information and Computer Science Department, University of California, Irvine, Irvine, CA, 2001.

[16] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[17] I. R. Cruz-Medina, T. P. Hettmansperger, and H. Thomas. Semiparametric Mixture Models and Repeated Measures: The Multinomial Cut Point Model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 53(3):463–474, 06 2004.

[18] Chen Dan, Liu Leqi, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. The sample complexity of semi-supervised learning with nonparametric mixture models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[19] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

[20] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. *arXiv e-prints*, page arXiv:1611.02648, November 2016.

[21] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1141–1148. MIT Press, 2004.

[22] Ruting Duan and Chunping Li. An adaptive dirichlet multinomial mixture model for short text streaming clustering. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 49–55, 2018.

[23] Ryan Elmore and Shaoli Wang. Identifiability and estimation in finite mixture models with multinomial components. Technical Report 03–04, Pennsylvania State University, Department of Statistics, 2003.

[24] Ryan T Elmore, Thomas P Hettmansperger, and Hoben Thomas. Estimating component cumulative distribution functions in finite mixture models. *Communications in Statistics-Theory and Methods*, 33(9):2075–2086, 2004.

[25] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C Holmes. Multi-facet clustering variational autoencoders. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8676–8690. Curran Associates, Inc., 2021.

[26] Gerald B. Folland. *Real analysis: modern techniques and their applications*. Pure and applied mathematics. Wiley, 1999.

[27] Joshua G. Harrison, W. John Calder, Vivaswat Shastry, and C. Alex Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources*, 20(2):481–497, 2020.

[28] T. P. Hettmansperger and Hoben Thomas. Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):811–825, 2000.

[29] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS ONE*, 7(2):1–15, 02 2012.

[30] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1965–1972. AAAI Press, 2017.

[31] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[32] Murray Jorgensen. Using multinomial mixture models to cluster internet traffic. *Australian & New Zealand Journal of Statistics*, 46(2):205–218, 2004.

[33] R.V. Kadison and J.R. Ringrose. *Fundamentals of the theory of operator algebras. V1: Elementary theory*. Pure and Applied Mathematics. Elsevier Science, 1983.

[34] O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer New York, 2002.

[35] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. 2005.

[36] Nikos Kargas and Nicholas D. Sidiropoulos. Learning mixtures of smooth product distributions: Identifiability and algorithm. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 388–396. PMLR, 16–18 Apr 2019.

[37] Nikos Kargas, Nicholas D. Sidiropoulos, and Xiao Fu. Tensors, learning, and "kolmogorov extension" for finite-alphabet random vectors. *IEEE Transactions on Signal Processing*, 66(18):4854–4868, 2018.

[38] Byung Soo Kim. *Studies of multinomial mixture models*. PhD thesis, The University of North Carolina at Chapel Hill, 1984.

[39] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15687–15701. Curran Associates, Inc., 2022.

[40] Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95 – 138, 1977.

[41] Michael Levine, David R Hunter, and Didier Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416, 2011.

[42] Minqiang Li and Liang Zhang. Multinomial mixture model with feature selection for text clustering. *Knowledge-Based Systems*, 21(7):704–708, 2008.

[43] Jocelyn Mazarura and Alta de Waal. A comparison of the performance of latent dirichlet allocation and the dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6, 2016.

[44] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[45] Krikamol Muandet and Bernhard Schölkopf. One-class support measure machines for group anomaly detection. *CoRR*, abs/1303.0309, 2013.

[46] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2):103–134, May 2000.

[47] Maria Osmala, Gökçen Eraslan, and Harri Lähdesmäki. ChromDMM: a Dirichlet-multinomial mixture model for clustering heterogeneous epigenetic data. *Bioinformatics*, 38(16):3863–3870, 07 2022.

[48] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry A. Wasserman. Distribution-free distribution regression. In *AISTATS*, volume 31 of *JMLR Proceedings*, pages 507–515. JMLR.org, 2013.

[49] J. Portela. Clustering discrete data through the multinomial mixture model. *Communications in Statistics - Theory and Methods*, 37(20):3250–3263, 2008.

[50] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 207–224, New York, NY, USA, 2014. ACM.

[51] John A. Rhodes. A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and its Applications*, 432:1818–1824, 2009.

[52] Loïs Rigouste, Olivier Cappé, and François Yvon. Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing & Management*, 43(5):1260–1280, 2007. Patent Processing.

[53] Alexander Ritchie, Robert A Vandermeulen, and Clayton Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11676–11686. Curran Associates, Inc., 2020.

[54] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 792–799, New York, NY, USA, 2005. ACM.

[55] Nicholas D. Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14(3):229–239, 2000.

[56] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–640–II–648. JMLR.org, 2014.

[57] Le Song and Bo Dai. Robust low rank kernel embeddings of multivariate distributions. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[58] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

[59] Wai Ming Tai and Bryon Aragam. Tight bounds on the hardness of learning simple nonparametric mixtures. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2849–2849. PMLR, 2023.

[60] Henry Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34(4):1265–1269, 12 1963.

[61] Jakub Tomczak and Max Welling. Vae with a vampprior. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223. PMLR, 09–11 Apr 2018.

[62] Robert A. Vandermeulen. Improving Nonparametric Density Estimation with Tensor Decompositions. *arXiv e-prints*, page arXiv:2010.02425, October 2020.

[63] Robert A. Vandermeulen. Sample Complexity Using Infinite Multiview Models. *arXiv e-prints*, page arXiv:2302.04292, February 2023.

[64] Robert A Vandermeulen and Antoine Ledent. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12180–12193. Curran Associates, Inc., 2021.

[65] Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric mixture models. *Ann. Statist.*, 47(5):2704–2733, 10 2019.

[66] Leena C. Vankadara, Sebastian Bordt, Ulrike von Luxburg, and Debarghya Ghoshdastidar. Recovery guarantees for kernel-based clustering under non-parametric mixture models. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3817–3825. PMLR, 13–15 Apr 2021.

[67] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6448–6457, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[68] Yun Wei and XuanLong Nguyen. Convergence of de Finetti's mixing measure in latent structure models for observed exchangeable sequences. *The Annals of Statistics*, 50(4):1859 – 1889, 2022.
[69] Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39(1):209–214, 02 1968.
[70] S Yantis, D E Meyer, and J E Smith. Analyses of multinomial mixture distributions: new tests for stochastic models of cognition and action. *Psychol Bull*, 110(2):350–374, September 1991.
[71] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 233–242, New York, NY, USA, 2014. Association for Computing Machinery.

Robert A. Vandermeulen received his PhD in EECS from the University of Michigan and is currently a researcher at the Berlin Institute for the Foundations of Learning and Data. His research interests include deep learning and nonparametric estimation.

René Saitenmacher received his B.Sc. in mathematics from Freie Universität Berlin and his M.Sc. in computer science from Technische Universität Berlin. He is currently a doctoral student at the Weierstrass Institute for Applied Analysis and Stochastics in Berlin, Germany. His research interests include optimization, nonparametric estimation and topics in theoretical machine learning.