The Annals of Statistics
2019, Vol. 47, No. 5, 2704–2733
https://doi.org/10.1214/18-AOS1762

AN OPERATOR THEORETIC APPROACH TO NONPARAMETRIC MIXTURE MODELS[1]

BY ROBERT A. VANDERMEULEN AND CLAYTON D. SCOTT

Technische Universität Kaiserslautern and University of Michigan

When estimating finite mixture models, it is common to make assumptions on the mixture components, such as parametric assumptions. In this work, we make no distributional assumptions on the mixture components and instead assume that observations from the mixture model are grouped, such that observations in the same group are known to be drawn from the same mixture component. We precisely characterize the number of observations n per group needed for the mixture model to be identifiable, as a function of the number m of mixture components. In addition to our assumption-free analysis, we also study the settings where the mixture components are either linearly independent or jointly irreducible. Furthermore, our analysis considers two kinds of identifiability, where the mixture model is the simplest one explaining the data, and where it is the only one. As an application of these results, we precisely characterize identifiability of multinomial mixture models. Our analysis relies on an operator-theoretic framework that associates mixture models in the grouped-sample setting with certain infinite-dimensional tensors. Based on this framework, we introduce a general spectral algorithm for recovering the mixture components.

1. Introduction. A finite mixture model $\mathscr{P}$ is a probability measure over a space of probability measures where $\mathscr{P}(\{\mu_i\}) = w_i > 0$ for some finite collection of probability measures $\mu_1, \ldots, \mu_m$ and $\sum_{i=1}^m w_i = 1$. A realization from this mixture model first randomly selects some mixture component $\mu \sim \mathscr{P}$ and then draws from μ. Mixture models have seen extensive use in statistics and machine learning.

A central theoretical question concerning mixture models is that of identifiability. A mixture model is said to be *identifiable* if there is no other mixture model that defines the same distribution over the observed data. Classically, mixture models were concerned with the case where the observed data $X_1, X_2, \ldots$ are i.i.d. with X_i distributed according to some unobserved random measure μ_i with $\mu_i \overset{\text{i.i.d.}}{\sim} \mathscr{P}$. This situation is equivalent to $X_i \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^m w_j \mu_j$. If we impose no restrictions on

Received October 2016; revised March 2018.

[1]Supported by NSF Grant 1422157, German Research Foundation (DFG) award KL 2698/21, and Federal Ministry of Science and Education (BMBF) award 031B0187B.

MSC2010 subject classifications. Primary 62E10; secondary 62G05.

Key words and phrases. Mixture model, nonparametric mixture, identifiability, tensor factorization, multinomial mixture, topic model, joint irreducibility.

the mixture components $\mu_1, \ldots, \mu_m$, one could easily concoct many choices of μ_j and w_j which yield an identical distribution on X_i. Because of this, most previous work on identifiability assumes some sort of structure on $\mu_1, \ldots, \mu_m$, such as Gaussianity [3, 9, 26]. In this work, we consider an alternative scenario where we make no assumptions on $\mu_1, \ldots, \mu_m$ and instead have access to groups of samples that are known to come from the same component. We will call these groups of samples "random groups." Mathematically, a random group is a random element $\mathbf{X}_i$ where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n})$ with $X_{i,1}, \ldots, X_{i,n} \overset{\text{i.i.d.}}{\sim} \mu_i$ and $\mu_i \overset{\text{i.i.d.}}{\sim} \mathscr{P}$.

In this setting, identifiability is now concerned with the distribution over $\mathbf{X}_i$ and the value of n, the number of samples in each random group. We call a mixture of measures $\mathscr{P}$ *n-identifiable* if it is the *simplest* mixture model (in terms of number of mixture components) that yields the observed distribution on $\mathbf{X}_i$. We also introduce a concept which is stronger than identifiability. We call $\mathscr{P}$ *n-determined* if it is the *only* mixture model that yields the observed distribution on $\mathbf{X}_i$.

In this paper, we show that every mixture model with m components is $(2m-1)$-identifiable and $2m$-determined. Furthermore, we show that any mixture model with linearly independent components is 3-identifiable and 4-determined, and any mixture model with jointly irreducible components is 2-determined. These results, presented in Section 4, hold for any mixture model over any space and cannot be improved. The operator theoretic framework underlying our analysis is presented in Section 5, and selected proofs of our main results appear in Section 6, with the rest appearing in the Supplementary Material [25]. In Section 7, we apply our main results to demonstrate some new and old results on the identifiability of multinomial mixture models. Section 8 describes a spectral algorithm for the recovery of the mixture components and weights, and experimental results on simulated data are presented in Section 9. Related work, the problem formulation and a concluding discussion are offered in Sections 2, 3 and 10, respectively.

To keep the paper length reasonable, many of the proofs have been omitted and can be found in the Supplementary Material [25]. The Supplementary Material [25] also contains an in-depth description of the application of our spectral algorithm to categorical data (including a consistency proof) and additional technical details regarding the experiments in Section 9.

2. Previous work. In classical mixture model theory, identifiability is achieved by making assumptions about the mixture components. Some assumptions which yield identifiability are Gaussian or binomial mixture components [9, 24]. If one makes no assumptions on the mixture components, then one must leverage some other type of structure in order to achieve identifiability. An example of such structure exists in the context of multiview models. In a multiview model, samples have the form $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n})$ and the distribution of $\mathbf{X}_i$ is defined by $\sum_{i=1}^m w_i \prod_{j=1}^n \mu_i^j$. In [1], it was shown that if μ_i^j are probability distributions on $\mathbb{R}$ with $\mu_1^j, \ldots, \mu_m^j$ linearly independent for all j and $n \geq 3$, then the model is

identifiable. In [7], the authors perform a *smoothed analysis* of tensor decompositions. They demonstrate that, with high probability, a tensor's components are both identifiable and can be recovered using a polynomial time algorithm, provided the component dimensionality is sufficiently high. In that paper, the authors go on to apply the result to multiview models, demonstrating bounds on identifiability.

The setting which we investigate is a special case of the multiview model where $\mu_i^j = \mu_i^{j'}$ for all i, j, j'. If the sample space of the μ_i is finite, then this problem is exactly the topic modeling problem with a finite number of topics and one topic for each document. In topic modeling, each μ_i is a "topic" and the sample space is a finite collection of words. This setting is well studied and it has been shown that one can recover the true topics provided certain assumptions on the topics are satisfied [1, 2, 4, 5]. This problem was studied for arbitrary topics in [22]. In this paper, the authors introduce an algorithm that recovers any mixture of m topics provided $2m - 1$ words per document. They also show, in a result analogous to our own, that this $2m - 1$ value cannot be improved. Our proof techniques are quite different than those used in [22], hold for arbitrary sample spaces and are less complex.

In Lemma 7.1, we show that, when restricted to finite sample spaces, the grouped sample setting introduced in this paper is equivalent to a multinomial mixture model. Fundamental bounds on the identifiability of multinomial mixture models can be found in [12, 17]. We will reproduce these results (and develop some new results) using techniques developed in this paper. Additional connections to previous work are given later.

3. Problem setup. We treat this problem in a general setting. For any measurable space, we define δ_x as the Dirac measure at x. For Υ a set, σ-algebra, or measure, we denote $\Upsilon^{\times a}$ to be the standard a-fold product associated with that object. Let $\mathbb{N}$ be the set of integers greater than or equal to zero and $\mathbb{N}_+$ be the integers strictly greater than 0. For $k \in \mathbb{N}_+$, we define $[k]$ to be those elements in $\mathbb{N}_+$ which are less than or equal to k. Let Ω be a set containing more than one element. This set is the sample space of our data. Let $\mathcal{F}$ be a σ-algebra over Ω. Assume $\mathcal{F} \neq \{\varnothing, \Omega\}$, that is, $\mathcal{F}$ contains nontrivial events. We denote the space of probability measures over a measurable space $(\Psi, \mathcal{G})$ as $\mathcal{D}(\Psi, \mathcal{G})$. The space $\mathcal{D}(\Omega, \mathcal{F})$ will be shortened to $\mathcal{D}$ for brevity. We equip $\mathcal{D}$ with the σ-algebra $2^{\mathcal{D}}$ so that each Dirac measure over $\mathcal{D}$ is unique. Define $\Delta(\mathcal{D}) \triangleq \text{span}(\{\delta_x : x \in \mathcal{D}\})$. This is the ambient space where our mixtures of probability measures live. Let $\mathscr{P} = \sum_{i=1}^m w_i \delta_{\mu_i}$ be a probability measure in $\Delta(\mathcal{D})$. Let $\mu \sim \mathscr{P}$ and $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$ and denote $\mathbf{X} = (X_1, \ldots, X_n)$. Here, $\mathbf{X}$ is a random group sample, which was described in the Introduction.

We now derive the probability distribution of $\mathbf{X}$. Let $A \in \mathcal{F}^{\times n}$. Letting $\mathbb{P}$ reflect both the draw of $\mu \sim \mathscr{P}$ and $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$, we have

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{i=1}^{m} \mathbb{P}(\mathbf{X} \in A | \mu = \mu_i)\mathbb{P}(\mu = \mu_i) \tag{1}$$

$$= \sum_{i=1}^{m} w_i \mu_i^{\times n}(A). \tag{2}$$

The second equality follows from Lemma 3.10 in [16]. So the probability distribution of $\mathbf{X}$ is

$$\sum_{i=1}^{m} w_i \mu_i^{\times n}. \tag{3}$$

We want to view the probability distribution of $\mathbf{X}$ as a function of $\mathscr{P}$ in a mathematically rigorous way, which requires a bit of technical buildup. Let $\mathscr{Q} \in \Delta(\mathcal{D})$. From the definition of $\Delta(\mathcal{D})$, it follows that $\mathscr{Q}$ admits a representation

$$\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i}.$$

From the well-ordering principle, there must exist some representation with minimal r and we define this r as the *order* of $\mathscr{Q}$. We can show that the minimal representation of any $\mathscr{Q} \in \Delta(\mathcal{D})$ is unique up to permutation of its indices.

LEMMA 3.1. *Let $\mathscr{Q} \in \Delta(\mathcal{D})$ and admit minimal representations $\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i} = \sum_{j=1}^{r} \alpha'_j \delta_{\nu'_j}$. There exists some permutation $\psi : [r] \to [r]$ such that $\nu_{\psi(i)} = \nu'_i$ and $\alpha_{\psi(i)} = \alpha'_i$ for all i.*

Proofs of most of the lemmas in this paper are omitted and can be found in the Supplementary Material [25]. The only lemma proved in this paper is Lemma 6.5 since it is nontrivial and quite crucial for showing some of our bounds are tight.

Henceforth, when we define an element of $\Delta(\mathcal{D})$ with a summation, we will assume that the summation is a minimal representation.

DEFINITION 3.1. We call $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ a *mixture of measures* if it is a probability measure in $\Delta(\mathcal{D})$. The elements $\mu_1, \ldots, \mu_m$, are called *mixture components*.

Any minimal representation of a mixture of measures $\mathscr{P}$ with m components satisfies $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ with $w_i > 0$ for all i and $\sum_{i=1}^{m} w_i = 1$. Hence any mixture of measures is a convex combination of Dirac measures at elements in $\mathcal{D}$.

For a measurable space $(\Psi, \mathcal{G})$, we define $\mathscr{M}(\Psi, \mathcal{G})$ as the space of all finite signed measures over $(\Psi, \mathcal{G})$. We can now introduce the operator $V_n : \Delta(\mathcal{D}) \to$

$\mathscr{M}(\Omega^{\times n}, \mathcal{F}^{\times n})$. For a minimal representation $\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i}$, we define V_n, with $n \in \mathbb{N}_+$, as

$$V_n(\mathscr{Q}) = \sum_{i=1}^{r} \alpha_i \nu_i^{\times n}. \tag{4}$$

This mapping is well defined as a consequence of Lemma 3.1. From this definition, we have that $V_n(\mathscr{P})$ is simply the distribution of $\mathbf{X}$ which we derived earlier. In the following definitions, two mixtures of measures are considered equal if they define the same measure.

DEFINITION 3.2. We call a mixture of measures, $\mathscr{P}$, *n-identifiable* if there does not exist a different mixture of measures $\mathscr{P}'$, with order no greater than the order of $\mathscr{P}$, such that $V_n(\mathscr{P}) = V_n(\mathscr{P}')$.

DEFINITION 3.3. We call a mixture of measures, $\mathscr{P}$, *n-determined* if there exists no other mixture of measures $\mathscr{P}'$ such that $V_n(\mathscr{P}) = V_n(\mathscr{P}')$.

Definitions 3.2 and 3.3 are central objects of interest in this paper. Given a mixture of measures, $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ then $V_n(\mathscr{P})$ is equal to $\sum_{i=1}^{m} w_i \mu_i^{\times n}$, the measure from which $\mathbf{X}$ is drawn. If $\mathscr{P}$ is not n-identifiable, then we know that there exists a different mixture of measures that is no more complex (in terms of number of mixture components) than $\mathscr{P}$ which induces the same distribution on $\mathbf{X}$. Practically speaking, this means we need more samples in each random group $\mathbf{X}$ in order for the full richness of $\mathscr{P}$ to be manifested in $\mathbf{X}$. A stronger version of n-identifiability is n-determinedness where we enforce the requirement that our mixture of measures be the *only* mixture of measures (of any order) that admits the distribution on $\mathbf{X}$.

A quick note on terminology. We use the term "mixture of measures" rather than "mixture model" to emphasize that a mixture of measures should be interpreted a bit differently than a typical mixture model. A "mixture model" connotes a probability measure on the sample space of observed data Ω, whereas a "mixture of measures" connotes a probability measure on the sample space of the unobserved latent measures $\mathcal{D}$.

4. Main results. The first result is a bound on the n-identifiability of all mixtures of measures with m or fewer components. This bound cannot be uniformly improved.

THEOREM 4.1. *Let* $(\Omega, \mathcal{F})$ *be a measurable space. Mixtures of measures with* m *components are* $(2m-1)$*-identifiable.*

THEOREM 4.2. *Let* $(\Omega, \mathcal{F})$ *be a measurable space with* $\mathcal{F} \neq \{\varnothing, \Omega\}$. *For all* $m \geq 2$, *there exists a mixture of measures with* m *components that is not* $(2m-2)$*-identifiable.*

We mention again that the previous two theorems had been previously found in [22] for finite sample spaces, using techniques different from our own. To be explicit, a "finite sample space" in our problem setting is the assumption that $|\Omega| < \infty$ and $\mathcal{F} = 2^{\Omega}$, which implies that the mixture components are categorical distributions. The following lemmas convey the unsurprising fact that n-identifiability is, in some sense, monotonic.

LEMMA 4.1. *If a mixture of measures is n-identifiable, then it is q-identifiable for all $q > n$.*

LEMMA 4.2. *If a mixture of measures is not n-identifiable, then it is not q-identifiable for any $q < n$.*

Viewed alternatively, these results say that $n = 2m - 1$ is the smallest value for which V_n is injective over the set of mixtures of measures with m or fewer components.

We also present an analogous bound for n-determinedness. This bound also cannot be improved.

THEOREM 4.3. *Let $(\Omega, \mathcal{F})$ be a measurable space. Mixtures of measures with m components are $2m$-determined.*

THEOREM 4.4. *Let $(\Omega, \mathcal{F})$ be a measurable space with $\mathcal{F} \neq \{\varnothing, \Omega\}$. For all m, there exists a mixture of measures with m components that is not $(2m-1)$-determined.*

Again n-determinedness is monotonic in the number of samples per group.

LEMMA 4.3. *If a mixture of measures is n-determined, then it is q-determined for all $q > n$.*

LEMMA 4.4. *If a mixture of measures is not n-determined, then it is not q-determined for any $q < n$.*

This collection of results can be interpreted in an alternative way. Consider some pair of mixtures of measures $\mathscr{P}$, $\mathscr{P}'$. If $n \geq 2m$ and either mixture of measures is of order m or less, then $V_n(\mathscr{P}) = V_n(\mathscr{P}')$ implies $\mathscr{P} = \mathscr{P}'$. Furthermore, $n = 2m$ is the smallest value of n for which the previous statement is true for all pairs of mixtures of measures.

Our definitions of n-identifiability, n-determinedness and their relation to previous works on identifiability deserve a bit of discussion. Some previous works on identifiability contain results related to what we call "identifiability" and others contain results related what we call "determinedness." Both of these are simply

called "identifiability" in these works. For example, in [26] it is shown that different finite mixtures of multivariate Gaussian distributions will always yield different distributions, a result which we could call "determinedness." Alternatively, in [24] it is demonstrated that mixtures of binomial distributions, with a fixed number of trials n for every mixture component, are identifiable provided we only consider mixtures with m mixture components and $n \geq 2m - 1$. In this result, allowing for more mixture components may destroy identifiability, and thus this is what *we* would call an "identifiability" result. The fact that the value $2m - 1$ occurs in both the previous binomial mixture model result and Theorem 4.1 is not a coincidence. We will demonstrate a new determinedness result for multinomial mixtures models later in the paper, under the assumption that $n \geq 2m$. We will prove these results using Theorems 4.1 and 4.3. To our knowledge, our work is the first to consider both identifiability and determinedness.

Finally, we also include results that are analogous to previously shown results for the finite sample space setting. We note that our proof techniques are markedly different than the previous proofs for the finite sample space case.

THEOREM 4.5. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are linearly independent, then $\mathscr{P}$ is* 3*-identifiable.*

This bound is tight as a consequence of Theorem 4.2 with $m = 2$ since any pair of distinct measures must be linearly independent.

A version of this theorem was first proven in [1] by making use of Kruskal's theorem [18]. Kruskal's theorem demonstrates that order 3 tensors over $\mathbb{R}^d$ admit unique decompositions (up to scaling and permutation) given certain linear independence assumptions. The linear independence assumption in Theorem 4.5 is stronger than that contained in Kruskal's theorem, and thus yields a simple proof which does not invoke Kruskal's theorem. An efficient algorithm for recovering linearly independent mixture components for finite sample spaces with 3 samples per random group is described in [2]. Interestingly, with one more sample per group, these mixtures of measures become determined.

THEOREM 4.6. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are linearly independent, then $\mathscr{P}$ is* 4*-determined.*

This bound is tight as a result of Theorem 4.4 with $m = 2$.

Our final result is related to the "separability condition" found in [11]. The separability condition in the finite sample space setting requires that, for each mixture component μ_i, there exists $B_i \in \mathcal{F}$ such that $\mu_i(B_i) > 0$ and $\mu_j(B_i) = 0$ for all $i \neq j$. There exists a generalization of the separability condition, known as *joint irreducibility*.

DEFINITION 4.1. A collection of probability measures $\mu_1, \ldots, \mu_m$ are said to be *jointly irreducible* if $\sum_{i=1}^{m} w_i \mu_i$ being a probability measure implies $w_i \geq 0$.

In other words, any probability measure in the span of $\mu_1, \ldots, \mu_m$ must be a convex combination of those measures. It was shown in [8] that separability implies joint irreducibility, but not vice versa. In that paper, it was also shown that joint irreducibility implies linear independence, but the converse does not hold.

THEOREM 4.7. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are jointly irreducible, then $\mathscr{P}$ is 2-determined.*

A straightforward consequence of the corollary of Theorem 1 in [11] is that any mixture of measures on a finite sample space with jointly irreducible components is 2-identifiable. The result in [11] is concerned with the uniqueness of nonnegative matrix factorizations and Theorem 4.7, when applied to a finite sample space, can be posed as a special case of the result in [11]. In the context of nonnegative matrix factorization, the result in [11] is significantly more general than our result. In another sense, our result is more general since it applies to spaces where joint irreducibility and the separability condition are not equivalent. Furthermore, [11] only implies that the mixture of measures in Theorem 4.7 are identifiable. The determinedness result is, as far as we know, totally new. It is worth mentioning that, in finite sample spaces, the separability condition assumption yields efficient algorithms for nonnegative matrix factorization [4, 5], whereas we are not aware of analogous algorithms which are applicable to the more general joint irreducibility setting.

5. Tensor products of Hilbert spaces. Our proofs will rely heavily on the geometry of tensor products of Hilbert spaces which we will introduce in this section.

5.1. *Overview of tensor products.* First, we introduce tensor products of Hilbert spaces. Instead of a rigorous primer to the subject, we will simply state some basic facts about tensor products of Hilbert spaces, and hopefully instill some intuition for the uninitiated by way of example. A thorough treatment of tensor products of Hilbert spaces can be found in [15].

Let H and H' be Hilbert spaces. From these two Hilbert spaces, the "simple tensors" are elements of the form $h \otimes h'$ with $h \in H$ and $h' \in H'$. We can define an inner product on the simple tensors by setting

$$\langle h_1 \otimes h_1', h_2 \otimes h_2' \rangle = \langle h_1, h_2 \rangle \langle h_1', h_2' \rangle. \tag{5}$$

Let H_0 be the inner product space spanned by the simple tensors. The tensor product of H and H' is the completion of H_0 and is denoted $H \otimes H'$. To avoid potential confusion, we note that the notation just described is standard in operator theory literature. In some literature, our definition of H_0 is denoted as $H \otimes H'$ and our definition of $H \otimes H'$ is denoted $H \widehat{\otimes} H'$.

As an illustrative example, we consider the tensor product $L^2(\mathbb{R}) \otimes L^2(\mathbb{R})$. It can be shown that there exists an isomorphism between $L^2(\mathbb{R}) \otimes L^2(\mathbb{R})$ and $L^2(\mathbb{R}^2)$ that maps the simple tensors to separable functions [15], $f \otimes f' \mapsto f(\cdot) f'(\cdot)$. We can demonstrate this isomorphism with a simple example. Let $f, g, f', g' \in L^2(\mathbb{R})$. Taking the $L^2(\mathbb{R}^2)$ inner product of $f(\cdot) f'(\cdot)$ and $g(\cdot) g'(\cdot)$ gives us

$$\int \int \big(f(x) f'(y)\big)\big(g(x) g'(y)\big)\, dx\, dy = \int f(x) g(x)\, dx \int f'(y) g'(y)\, dy \tag{6}$$

$$= \langle f, g \rangle \langle f', g' \rangle \tag{7}$$

$$= \langle f \otimes f', g \otimes g' \rangle. \tag{8}$$

Beyond tensor product, we will need to define tensor power. To begin, we will first show that tensor products are, in a certain sense, associative. Let H_1, H_2, H_3 be Hilbert spaces. Proposition 2.6.5 in [15] states that there is a unique unitary operator, $U : (H_1 \otimes H_2) \otimes H_3 \to H_1 \otimes (H_2 \otimes H_3)$, that satisfies the following for all $h_1 \in H_1, h_2 \in H_2, h_3 \in H_3$:

$$U\big((h_1 \otimes h_2) \otimes h_3\big) = h_1 \otimes (h_2 \otimes h_3). \tag{9}$$

This implies that for any collection of Hilbert spaces, $H_1, \ldots, H_n$, the Hilbert space $H_1 \otimes \cdots \otimes H_n$ is defined unambiguously regardless of how we decide to associate the products. In the space $H_1 \otimes \cdots \otimes H_n$, we define a *simple tensor* as a vector of the form $h_1 \otimes \cdots \otimes h_n$ with $h_i \in H_i$. In [15], it is shown that $H_1 \otimes \cdots \otimes H_n$ is the closure of the span of these simple tensors. To conclude this primer on tensor products, we introduce the following notation. For a Hilbert space H, we denote $H^{\otimes n} = \underbrace{H \otimes H \otimes \cdots \otimes H}_{n \text{ times}}$ and for $h \in H$, $h^{\otimes n} = \underbrace{h \otimes h \otimes \cdots \otimes h}_{n \text{ times}}$.

5.2. *Tensor rank.* A tool we will use frequently in our proofs is *tensor rank*, which is similar to matrix rank.

DEFINITION 5.1. Let $h \in H^{\otimes n}$ where H is a Hilbert space. The *rank* of h is the smallest natural number r such that $h = \sum_{i=1}^{r} h_i$ where h_i are simple tensors.

In an infinite dimensional Hilbert space, it is possible for a tensor to have infinite rank. We will only be concerned with finite rank tensors.

5.3. *Some results for tensor product spaces.* We present some technical results concerning tensor product spaces that will be useful for the rest of the paper. These lemmas are similar to or are straightforward extensions of previous results which we needed to modify for our particular purposes. The following lemma is used in the proof of Lemma 5.2 (Supplementary Material [25]) and Lemma 6.5.

LEMMA 5.1. *Let $H_1, \dots, H_n$, $H_1', \dots, H_n'$ be a collection of Hilbert spaces and $U_1, \dots, U_n$ a collection of unitary operators with $U_i : H_i \to H_i'$ for all i. There exists a unitary operator $U : H_1 \otimes \cdots \otimes H_n \to H_1' \otimes \cdots \otimes H_n'$ satisfying $U(h_1 \otimes \cdots \otimes h_n) = U_1(h_1) \otimes \cdots \otimes U_n(h_n)$ for all $h_1 \in H_1, \dots, h_n \in H_n$.*

Let $(\Psi, \mathcal{G}, \gamma)$ be a σ-finite measure space. We have the following lemma that connects tensor power of a L^2 space to the L^2 space of the product measure.

LEMMA 5.2. *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 $f_1, \dots, f_n \in L^2(\Psi, \mathcal{G}, \gamma)$,*

$$U(f_1 \otimes \cdots \otimes f_n) = f_1(\cdot) \cdots f_n(\cdot). \tag{10}$$

A statement of the following lemma for $\mathbb{R}^d$ can be found in [10]. We present our own proof for the Hilbert space setting in the Supplementary Material [25].

LEMMA 5.3. *Let $n > 1$ and let $h_1, \dots, h_n$ be elements of a Hilbert space such that no elements are zero and no pairs of elements are collinear. Then $h_1^{\otimes n-1}, \dots, h_n^{\otimes n-1}$ are linearly independent.*

The following lemma is a Hilbert space version of a well-known property for positive semidefinite matrices.

LEMMA 5.4. *Let $h_1, \dots, h_m$ be elements of a Hilbert space. The rank of $\sum_{i=1}^m h_i^{\otimes 2}$ is the dimension of* $\operatorname{span}(\{h_1, \dots, h_m\})$.

6. Proofs of theorems. With the tools developed in the previous sections, we will now prove a few, selected theorems. Due to space constraints, we only prove Theorems 4.1 to 4.4 in this document. These proofs give a good overview of the general techniques used to prove the other identifiability and determinedness results, which can be found in the Supplementary Material [25]. These theorems are proved for general measure spaces which introduces a fair amount of mathematical overhead. There are two basic components to these proofs, transforming the measure problem into and out of a tensor framework, and using the tensor framework as a means to manipulate these objects geometrically. For concreteness, it can be helpful to consider the situation where $\Omega = \{1, 2, \dots, d\}$, that is, a finite sample space. In this situation, a mixture component μ can be directly associated with a probability vector $p \in \mathbb{R}^d$ where $[p]_i = \mu(\{i\})$ and the tensor $p^{\otimes m}$ represents the density of m i.i.d. samples of μ:

$$\left[p^{\otimes m}\right]_{i_1,\dots,i_m} = [p]_{i_1}[p]_{i_2} \cdots [p]_{i_m} \tag{11}$$

$$= \mu(\{i_1\})\mu(\{i_2\}) \cdots \mu(\{i_m\}) \tag{12}$$

$$= \mu^{\times m}(\{(i_1, i_2, \dots, i_m)\}). \tag{13}$$

In essence, all of our results directly parallel this setting once transformed into the tensor space.

Before we begin our proofs, we need to introduce one additional piece of notation. For a function f on a domain $\mathcal{X}$, we define $f^{\times k}$ as simply the product of the function k times on the domain $\mathcal{X}^{\times k}$,

$$f^{\times k} = \underbrace{f(\cdot)\cdots f(\cdot)}_{k \text{ times}}. \tag{14}$$

For a set, σ-algebra, or measure the notation continues to denote the standard k-fold product.

In these proofs, we will be making extensive use of various L^2 spaces. These spaces will be equivalence classes of functions which are equal almost everywhere with respect to the measure associated with that space. When considering elements of these spaces, equality will always mean almost everywhere equality with respect to the measure associated with that space. When performing integrals or other manipulations of elements in L^2 spaces, we will be performing operations that do not depend on the representative of the equivalence class. The following lemma will be quite useful.

LEMMA 6.1. *Let $\gamma_1, \ldots, \gamma_m, \pi_1, \ldots, \pi_l$ be probability measures on a measurable space $(\Psi, \mathcal{G})$, $a_1, \ldots, a_m, b_1, \ldots, b_l \in \mathbb{R}$ and $n \in \mathbb{N}_+$. If*

$$\sum_{i=1}^{m} a_i \gamma_i^{\times n} = \sum_{j=1}^{l} b_j \pi_j^{\times n} \tag{15}$$

then for all $n' \in \mathbb{N}_+$ with $n' \leq n$ we have that

$$\sum_{i=1}^{m} a_i \gamma_i^{\times n'} = \sum_{j=1}^{l} b_j \pi_j^{\times n'}. \tag{16}$$

PROOF OF THEOREM 4.1. We proceed by contradiction. Suppose there exist $m, l \in \mathbb{N}_+$ with $l \leq m$ such that there two different mixtures of measures $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i} \neq \mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$, and

$$\sum_{i=1}^{m} a_i \mu_i^{\times 2m-1} = \sum_{j=1}^{l} b_j \nu_j^{\times 2m-1}. \tag{17}$$

Clearly, $m > 1$ otherwise we immediately arrive at a contradiction. By the well-ordering principle, there exists a minimal m such that the previous statement holds. For that minimal m, there exists a minimal l such that the previous statement holds. We will assume that the m and l are both minimal in this way. This assumption implies that $\mu_i \neq \nu_j$ for all i, j. To prove this, we will assume that there exists i, j such that $\mu_i = \nu_j$, and show that this assumption leads to a contradiction. Without

loss of generality, we will assume that $\mu_m = \nu_l$. We will consider the three cases where $a_m = b_l$, $a_m > b_l$ and $a_m < b_l$.

Case 1. If $a_m = b_l$, then we have that

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu^{\times 2m-1} \tag{18}$$

and from Lemma 6.1 we have

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2(m-1)-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu^{\times 2(m-1)-1}. \tag{19}$$

Setting $\mathscr{P} = \sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \delta_{\mu_i}$ and $\mathscr{P}' = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \delta_{\nu_j}$, it now follows that $V_{2(m-1)-1}(\mathscr{P}) = V_{2(m-1)-1}(\mathscr{P}')$ which contradicts the minimality of m.

Case 2. If $a_m > b_l$, then we have

$$\sum_{i=1}^{m-1} \frac{a_i}{1-b_l} \mu_i^{\times 2m-1} + \frac{a_m - b_l}{1-b_l} \mu_m^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu_j^{\times 2m-1} \tag{20}$$

which contradicts the minimality of l by an argument similar to that in Case 1.

Case 3. If $a_m < b_l$, we have that

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-a_m} \nu_j^{\times 2m-1} + \frac{b_l - a_m}{1-a_m} \nu_l^{\times 2m-1}. \tag{21}$$

Again, we will use arguments similar to the one used in Case 1. If $l = m$, then swapping the mixtures associated with m and l gives us a pair of mixtures of measures which violates the minimality of l. If $l < m$, then from Lemma 6.1 we have that

$$\begin{aligned} &\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2(m-1)-1} \\ &\quad = \sum_{j=1}^{l-1} \frac{b_j}{1-a_m} \nu_j^{\times 2(m-1)-1} + \frac{b_l - a_m}{1-a_m} \nu_l^{\times 2(m-1)-1}, \end{aligned} \tag{22}$$

which violates the minimality of m, thus completing Case 3.

We have now established that $\mu_i \neq \nu_j$, for all i, j. We will use the following lemma to embed the mixture components in a Hilbert space.

LEMMA 6.2. *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 $B \in \mathcal{G}$,*

$$\gamma_i(B) = \int_B f_i \, d\pi. \tag{23}$$

From Lemma 6.2, there exists a finite measure ξ and nonnegative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\mu_i(B) = \int_B p_i \, d\xi$ and $\nu_j(B) = \int_B q_j \, d\xi$ for all i, j. Clearly, no two of these functions are equal (in the ξ-almost everywhere sense). If one of the functions were a scalar multiple of another, for example, $p_1 = \alpha p_2$ for some $\alpha \neq 1$, it would imply

$$\mu_1(\Omega) = \int p_1 \, d\xi = \int \alpha p_2 \, d\xi = \alpha. \tag{24}$$

This is not true so no pair of these functions are collinear.

We can use the following lemma to extend this new representation to a product measure.

LEMMA 6.3. *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*

$$\gamma^{\times n}(B) = \int_B f^{\times n} \, d\pi^{\times n}. \tag{25}$$

Thus for any $R \in \mathcal{F}^{\times 2m-1}$ we have

$$\int_R \sum_{i=1}^{m} a_i \, p_i^{\times 2m-1} \, d\xi^{\times 2m-1} = \sum_{i=1}^{m} a_i \mu_i^{\times 2m-1}(R) \tag{26}$$

$$= \sum_{j=1}^{l} b_j \nu_j^{\times 2m-1}(R) \tag{27}$$

$$= \int_R \sum_{j=1}^{l} b_j q_j^{\times 2m-1} \, d\xi^{\times 2m-1}. \tag{28}$$

The following lemma is a well known result in real analysis (Proposition 2.23 in [13]), but it is worth mentioning explicitly.

LEMMA 6.4. *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 $A \in \mathcal{G}$, $\int_A f \, d\gamma = \int_A g \, d\gamma$.*

From this lemma, it follows that

$$\sum_{i=1}^{m} a_i \, p_i^{\times 2m-1} = \sum_{j=1}^{l} b_j q_j^{\times 2m-1}. \tag{29}$$

Applying the U^{-1} operator from Lemma 5.2 to the previous equation yields

$$\sum_{i=1}^{m} a_i \, p_i^{\otimes 2m-1} = \sum_{j=1}^{l} b_j q_j^{\otimes 2m-1}. \tag{30}$$

Since $l + m \leq 2m$, Lemma 5.3 states that

$$p_1^{\otimes 2m-1}, \ldots, p_m^{\otimes 2m-1}, q_1^{\otimes 2m-1}, \ldots, q_l^{\otimes 2m-1} \tag{31}$$

are all linearly independent, and thus $a_i = 0$ and $b_j = 0$ for all i, j, a contradiction. □

PROOF OF THEOREM 4.2. To prove this theorem, we will construct a pair of mixture of measures, $\mathscr{P} \neq \mathscr{P}'$ which both contain m components and satisfy $V_{2m-2}(\mathscr{P}) = V_{2m-2}(\mathscr{P}')$. From our definition of $(\Omega, \mathcal{F})$, we know there exists $F \in \mathcal{F}$ such that F and F^C are nonempty. Let $x \in F$ and $x' \in F^C$. It follows that δ_x and $\delta_{x'}$ are different probability measures on $(\Omega, \mathcal{F})$. The theorem follows from the next lemma. We will prove the lemma after the theorem proof.

LEMMA 6.5. *Let $(\Psi, \mathcal{G})$ be a measurable space and γ, γ' be distinct probability measures on that space. Let $\varepsilon_1, \ldots, \varepsilon_t$ be $t \geq 3$ distinct values in $[0, 1]$. Then there exist $\beta_1, \ldots, \beta_t$, a permutation $\sigma : [t] \to [t]$ and $l \in \mathbb{N}_+$ such that*

$$\begin{aligned} &\sum_{i=1}^{l} \beta_i \big(\varepsilon_{\sigma(i)}\gamma + (1 - \varepsilon_{\sigma(i)})\gamma'\big)^{\times t-2} \\ &\quad = \sum_{j=l+1}^{t} \beta_j \big(\varepsilon_{\sigma(j)}\gamma + (1 - \varepsilon_{\sigma(j)})\gamma'\big)^{\times t-2} \end{aligned} \tag{32}$$

where $\beta_i > 0$ for all i, $\sum_{i=1}^{l} \beta_i = \sum_{j=l+1}^{t} \beta_j = 1$, and $l, t - l \geq \lfloor \frac{t}{2} \rfloor$.

Let $\varepsilon_1, \ldots, \varepsilon_{2m} \in [0, 1]$ be distinct and let $\mu_i = \varepsilon_i \delta_x + (1 - \varepsilon_i)\delta_{x'}$ for $i \in [2m]$. From Lemma 6.5 with $t = 2m$, there exists a permutation $\sigma : [2m] \to [2m]$ and $\beta_1, \ldots, \beta_{2m}$ such that

$$\sum_{i=1}^{m} \beta_i \mu_{\sigma(i)}^{\times 2m-2} = \sum_{j=m+1}^{2m} \beta_j \mu_{\sigma(j)}^{\times 2m-2}, \tag{33}$$

with $\sum_{i=1}^{m} \beta_i = \sum_{j=m+1}^{2m} \beta_j = 1$ and $\beta_i > 0$ for all i.

If we let $\mathscr{P} = \sum_{i=1}^{m} \beta_i \delta_{\mu_{\sigma(i)}}$ and $\mathscr{P}' = \sum_{j=m+1}^{2m} \beta_j \delta_{\mu_{\sigma(j)}}$, we have that $V_{2m-2}(\mathscr{P}) = V_{2m-2}(\mathscr{P}')$ and $\mathscr{P} \neq \mathscr{P}'$ since $\mu_1, \ldots, \mu_{2m}$ are distinct. □

For the next proof, we will introduce some notation. For a tensor $U \in \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_l}$, we define $U_{i_1,\ldots,i_l}$ to be the entry in the $[i_1, \ldots, i_l]$ location of U.

PROOF OF LEMMA 6.5. From Lemma 6.2, there exists a finite measure π and nonnegative functions $f, f' \in L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ such that, for all $A \in \mathcal{G}$, $\gamma(A) = \int_A f \, d\pi$ and $\gamma'(A) = \int_A f' \, d\pi$.

Let H_2 be the Hilbert space associated with the subspace in $L^2(\Psi, \mathcal{G}, \pi)$ spanned by f and f'. Let $(f_i)_{i=1}^t$ be nonnegative functions in $L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ with $f_i = \varepsilon_i f + (1 - \varepsilon_i) f'$. Clearly, f_i is a pdf over π for all i and there are no pairs in this collection which are collinear. Since H_2 is isomorphic to $\mathbb{R}^2$ there exists a unitary operator $U : H_2 \to \mathbb{R}^2$. From Lemma 5.1, there exists a unitary operator $U_{t-2} : H_2^{\otimes t-2} \to \mathbb{R}^{2^{\otimes t-2}}$, with $U_{t-2}(h_1 \otimes \cdots \otimes h_{t-2}) = U(h_1) \otimes \cdots \otimes U(h_{t-2})$. Because U is unitary, it follows that

$$U_{t-2}\big(\operatorname{span}\big(\{h^{\otimes t-2} : h \in H_2\}\big)\big) = \operatorname{span}\big(\{x^{\otimes t-2} : x \in \mathbb{R}^2\}\big). \tag{34}$$

An order r tensor, $A_{i_1,\ldots,i_r}$, is *symmetric* if $A_{i_1,\ldots,i_r} = A_{i_{\psi(1)},\ldots,i_{\psi(r)}}$ for any $i_1, \ldots, i_r$ and permutation $\psi : [r] \to [r]$. A consequence of Lemma 4.2 in [10] is that $\operatorname{span}(\{x^{\otimes t-2} : x \in \mathbb{R}^2\}) \subset S^{t-2}(\mathbb{C}^2)$, the space of all symmetric order $t-2$ tensors over $\mathbb{C}^2$. Complex symmetric tensor spaces will always be viewed as a vector space over the complex numbers and real symmetric tensor spaces will be always be viewed as a vector space over the real numbers.

From Proposition 3.4 in [10], it follows that the dimension of $S^{t-2}(\mathbb{C}^2)$ is $\binom{2+t-2-1}{t-2} = t - 1$. From this, it follows that $\dim S^{t-2}(\mathbb{R}^2) \leq t - 1$, where $S^{t-2}(\mathbb{R}^2)$ is the space of all symmetric order $t-2$ tensors over $\mathbb{R}^2$. To see this, consider some set of linearly dependent tensors $x_1, \ldots, x_r \in S^{t-2}(\mathbb{C}^2)$ each containing only real valued entries, that is, the tensors are in $S^{t-2}(\mathbb{R}^2)$. Then it follows that there exists $c_1, \ldots, c_r \in \mathbb{C}$ such that

$$\sum_{i=1}^r c_i x_i = 0. \tag{35}$$

Let $\Re$ denote the real component when applied to an element of $\mathbb{C}$, and the real component applied entrywise when applied to a tensor. We have that

$$0 = \Re\left(\sum_{i=1}^r c_i x_i\right) = \sum_{i=1}^r \Re(c_i x_i) = \sum_{i=1}^r \Re(c_i) x_i. \tag{36}$$

Thus it follows that $x_1, \ldots, x_r$ are linearly dependent in $S^{t-2}(\mathbb{R}^2)$, and thus the dimensionality bound holds, $\dim S^{t-2}(\mathbb{R}^2) \leq t - 1$.

From this, we get that

$$\dim\big(\operatorname{span}\big(\{h^{\otimes t-2} : h \in H_2\}\big)\big) \leq t - 1. \tag{37}$$

The bound on the dimension of $\operatorname{span}(\{h^{\otimes t-2} : h \in H_2\})$ implies that $(f_i^{\otimes t-2})_{i=1}^t$ are linearly dependent. Conversely, Lemma 5.3 implies that removing a single vector from $(f_i^{\otimes t-2})_{i=1}^t$ yields a set of vectors which are linearly independent. It follows that there exists $(\alpha_i)_{i=1}^t$ with $\alpha_i \neq 0$ for all i and

$$\sum_{i=1}^t \alpha_i f_i^{\otimes t-2} = 0. \tag{38}$$

There exists a permutation $\sigma : [t] \to [t]$ such that $\alpha_{\sigma(i)} < 0$ for all $i \in [l]$ and $\alpha_{\sigma(j)} > 0$ for all $j > l$ with $l \leq \lfloor \frac{t}{2} \rfloor$ (ensuring that $l \leq \lfloor \frac{t}{2} \rfloor$ may also require multiplying (38) by -1). This σ appears in the lemma statement, but for the remainder of the proof we will simply assume without loss of generality that $\alpha_i < 0$ for $i \in [l]$ with $l \leq \lfloor \frac{t}{2} \rfloor$.

From this, we have

$$\sum_{i=1}^{l} -\alpha_i f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \alpha_j f_j^{\otimes t-2}. \tag{39}$$

From Lemma 5.2, we have

$$\sum_{i=1}^{l} -\alpha_i f_i^{\times t-2} = \sum_{j=l+1}^{t} \alpha_j f_j^{\times t-2} \tag{40}$$

and thus

$$\int \sum_{i=1}^{l} -\alpha_i f_i^{\times t-2} \, d\pi^{\times t-2} = \int \sum_{j=l+1}^{t} \alpha_j f_j^{\times t-2} \, d\pi^{\times t-2} \tag{41}$$

$$\Rightarrow \quad \sum_{i=1}^{l} -\alpha_i = \sum_{j=l+1}^{t} \alpha_j. \tag{42}$$

Let $r = \sum_{i=1}^{l} -\alpha_i$. We know $r > 0$ so dividing both sides of (39) by r gives us

$$\sum_{i=1}^{l} -\frac{\alpha_i}{r} f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \frac{\alpha_j}{r} f_j^{\otimes t-2} \tag{43}$$

where the left-hand and the right-hand side are convex combinations. Let $(\beta_i)_{i=1}^{t}$ be positive numbers with $\beta_i = \frac{-\alpha_i}{r}$ for $i \in [l]$ and $\beta_j = \frac{\alpha_j}{r}$ for $j \in [t] \setminus [l]$. This gives us

$$\sum_{i=1}^{l} \beta_i f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes t-2}. \tag{44}$$

We will now consider 3 cases for the value of t.

Case 1. If $t = 3$, then $l = 1$ and $l, t - l \geq \lfloor \frac{t}{2} \rfloor$ is satisfied.

Case 2. If t is divisible by two, then we can do the following:

$$\sum_{i=1}^{l} \beta_i f_i^{\otimes \frac{t}{2}-1} \otimes f_i^{\otimes \frac{t}{2}-1} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes \frac{t}{2}-1} \otimes f_j^{\otimes \frac{t}{2}-1}. \tag{45}$$

Consider the elements in the last equation as order two tensors in $L^2(\Psi, \mathcal{G}, \pi)^{\otimes \frac{t}{2}-1} \otimes L^2(\Psi, \mathcal{G}, \pi)^{\otimes \frac{t}{2}-1}$. From Lemma 5.3 and Lemma 5.4, we have that the

RHS of the previous equation has rank at least $\frac{t}{2}$ and since $l \leq \frac{t}{2}$ it follows that $l = \frac{t}{2}$. Again, we have that $l, t-l \geq \lfloor \frac{t}{2} \rfloor$.

Case 3. If t is greater than 3 and not divisible by 2, then we can apply Lemma 5.2 to get

$$\int_\Psi \sum_{i=1}^{l} \beta_i f_i^{\times t-3} f_i(x)\, d\pi(x) = \int_\Psi \sum_{j=l+1}^{t} \beta_j f_j^{\times t-3} f_j(y)\, d\pi(y) \tag{46}$$

$$\Rightarrow \quad \sum_{i=1}^{l} \beta_i f_i^{\times t-3} = \sum_{j=l+1}^{t} \beta_j f_j^{\times t-3}. \tag{47}$$

Applying Lemma 5.2 again, we get

$$\sum_{i=1}^{l} \beta_i f_i^{\otimes t-3} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes t-3} \tag{48}$$

$$\Rightarrow \quad \sum_{i=1}^{l} \beta_i f_i^{\otimes \frac{t-1}{2}-1} \otimes f_i^{\otimes \frac{t-1}{2}-1} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes \frac{t-1}{2}-1} \otimes f_j^{\otimes \frac{t-1}{2}-1}. \tag{49}$$

Recall that $\lfloor \frac{t}{2} \rfloor \geq l$ so we also have that

$$\left\lfloor \frac{t}{2} \right\rfloor - l \geq 0 \tag{50}$$

$$\Rightarrow \quad \frac{t}{2} - l \geq -\frac{1}{2} \tag{51}$$

$$\Rightarrow \quad t - l \geq \frac{t-1}{2}. \tag{52}$$

From Lemma 5.3 and Lemma 5.4, we have that the RHS of (49) has rank at least $\frac{t-1}{2}$, and thus $l \geq \frac{t-1}{2}$. From this, we have that $t-l, l \geq \lfloor \frac{t}{2} \rfloor$ once again, which completes Case 3.

So $l, t-l \geq \lfloor \frac{t}{2} \rfloor$ for any $t \geq 3$. Applying Lemma 5.2 to (44), we have

$$\sum_{i=1}^{l} \beta_i f_i^{\times t-2} = \sum_{j=l+1}^{t} \beta_j f_j^{\times t-2}. \tag{53}$$

From Lemma 6.3, we have

$$\sum_{i=1}^{l} \beta_i \left(\varepsilon_i \gamma + (1-\varepsilon_i)\gamma'\right)^{\times t-2} = \sum_{j=l+1}^{t} \beta_j \left(\varepsilon_j \gamma + (1-\varepsilon_j)\gamma'\right)^{\times t-2}. \tag{54}$$

□

PROOF OF THEOREM 4.3. Let $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ and $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$ be mixtures of measures such that $\mathscr{P}' \neq \mathscr{P}$. We will proceed by contradiction. Suppose that $\sum_{i=1}^{m} a_i \mu_i^{\times 2m} = \sum_{j=1}^{l} b_j \nu_j^{\times 2m}$. From Theorem 4.1, we know that $\mathscr{P}$ is

$2m-1$-identifiable and, therefore, $2m$-identifiable by Lemma 4.1. It follows that $l > m$. From Lemma 6.2, there exists a finite measure ξ and nonnegative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\mu_i(B) = \int_B p_i \, d\xi$ and $\nu_j(B) = \int_B q_j \, d\xi$ for all i, j. Using Lemmas 6.3 and 6.4, we have

$$\sum_{i=1}^{m} a_i p_i^{\times 2m} = \sum_{j=1}^{l} b_j q_j^{\times 2m}. \tag{55}$$

By Lemma 5.2, we have

$$\sum_{i=1}^{m} a_i p_i^{\otimes 2m} = \sum_{j=1}^{l} b_j q_j^{\otimes 2m} \tag{56}$$

and, therefore,

$$\sum_{i=1}^{m} a_i p_i^{\otimes m} \otimes p_i^{\otimes m} = \sum_{j=1}^{l} b_j q_j^{\otimes m} \otimes q_j^{\otimes m}. \tag{57}$$

Consider the elements in the last equation as tensors in $L^2(\Omega, \mathcal{F}, \xi)^{\otimes m} \otimes L^2(\Omega, \mathcal{F}, \xi)^{\otimes m}$. Since no pair of vectors in $p_1, \ldots, p_m$ are collinear, from Lemma 5.3 and Lemma 5.4 we know that the LHS has rank m. On the other hand, no pair of vectors $q_1, \ldots, q_l$ are collinear either, so Lemma 5.3 says that there is a subset of $\{q_1^{\otimes m}, \ldots, q_l^{\otimes m}\}$ which contains at least $m+1$ linearly independent elements. By Lemma 5.4, it follows that the RHS has rank at least $m+1$, a contradiction. □

PROOF OF THEOREM 4.4. To prove this theorem, we will construct a pair of mixture of measures, $\mathscr{P} \neq \mathscr{P}'$ which contain m and $m+1$ components, respectively, and satisfy $V_{2m-1}(\mathscr{P}) = V_{2m-1}(\mathscr{P}')$. From our definition of $(\Omega, \mathcal{F})$, we know there exists $F \in \mathcal{F}$ such that F, F^C are nonempty. Let $x \in F$ and $x' \in F^C$. It follows that δ_x and $\delta_{x'}$ are different probability measures on $(\Omega, \mathcal{F})$. Let $\varepsilon_1, \ldots, \varepsilon_{2m+1}$ be distinct values in $[0, 1]$. Applying Lemma 6.5 with $t = 2m+1$ and letting $\mu_i = \varepsilon_i \delta_x + (1 - \varepsilon_i)\delta_{x'}$, there exists a permutation $\sigma : [2m+1] \to [2m+1]$ and $\beta_1, \ldots, \beta_{2m+1}$, with $\beta_i > 0$ for all i and $\sum_{i=1}^{m} \beta_i = \sum_{j=m+1}^{2m+1} \beta_j = 1$, such that

$$\sum_{i=1}^{m} \beta_i \mu_{\sigma(i)}^{\times 2m-1} = \sum_{j=m+1}^{2m+1} \beta_j \mu_{\sigma(j)}^{\times 2m-1}. \tag{58}$$

If we let $\mathscr{P} = \sum_{i=1}^{m} \beta_i \delta_{\mu_{\sigma(i)}}$ and $\mathscr{P}' = \sum_{j=m+1}^{2m+1} \beta_j \delta_{\mu_{\sigma(j)}}$, then it follows that $V_{2m-1}(\mathscr{P}) = V_{2m-1}(\mathscr{P}')$. □

7. Identifiability and determinedness of mixtures of multinomial distributions. Using the previous results, we can show analogous identifiability and determinedness results for mixtures of multinomial distributions. The identifiability of mixtures of multinomial distributions was originally studied in [17] which contains a proof of Corollary 7.1 from this paper. An alternative proof of this corollary can be found in [12]. These results are analogous to identifiability results presented in this paper. Our proofs (see the Supplementary Material [25]) use techniques which are very different from those used in [12, 17]. Our techniques can also be used to prove a determinedness style result, Corollary 7.2, which we have not seen addressed elsewhere in the multinomial mixture model literature.

Central to the results in this section is Lemma 7.1 which establishes an equivalence between the grouped sample setting and multinomial mixture models. A sample from a multinomial distribution can be viewed as totalling the outcomes from an i.i.d. sampling of a categorical distribution. Consider some probability measure μ over a finite sample space and let $\mathbf{X} = (X_1, \ldots, X_m)$ be a collection of m i.i.d. samples from μ. Here, $\mathbf{X}$ has the form of what we would call a "random group." Because $\mathbf{X}$ contains *i.i.d.* sample, no useful statistical information is contained in the order of the samples. It follows that we can simply tally the number of results for each outcome and not lose any useful statistical information. Lemma 7.1 formalizes this intuition so that we can apply tools developed earlier in this paper to the multinomial mixture model setting.

Before stating our results, we must first introduce some definitions and notation. Any multinomial distribution is completely characterized by positive integers n and q and a probability vector in $\mathbb{R}^q$, $p = [p_1, \ldots, p_q]^T$.The value q represents the number of possible outcomes of a trial, p is the likelihood of each outcome on a trial and n is the number of trials. For whole numbers k, l, we define $C_{k,l} = \{x \in \mathbb{N}^{\times l} : \sum_{i=1}^{l} x_i = k\}$. These are vectors of the form $[x_1, \ldots, x_l]^T$ where $\sum_{i=1}^{l} x_i = k$. Using the values n and q above, the multinomial distribution is a probability measure over $C_{n,q}$. If Q is a multinomial distribution with parameters n, p, q as defined above, then its probability mass function is

$$Q(\{[x_1, \ldots, x_q]^T\}) = \frac{n!}{x_1! \cdots x_q!} p_1^{x_1} \cdots p_q^{x_q} \tag{59}$$

for $x \in C_{n,q}$. We will denote this measure as $Q_{n,p,q}$. Let

$$\mathscr{M}(n, q) \triangleq \{Q_{n,p,q} : p \text{ is a probability vector in } \mathbb{R}^q\}, \tag{60}$$

that is, the space of all multinomial distributions with n and q fixed.

At the heart of our multinomial mixture model, identifiability and determinedness results is the construction of a linear operator $T_{n,q}$ from $\operatorname{span}(\mathcal{D}(C_{n,q}, 2^{C_{n,q}}))$ to $\operatorname{span}(\mathcal{D}([q]^{\times n}, 2^{[q]^{\times n}}))$ and its use to show that nonidentifiable mixtures of multinomial distributions yield nonidentifiable mixtures of measures and nondetermined mixtures of multinomial distributions yields nondetermined mixtures of measures.

Since $C_{n,q}$ is a finite set, the vector space of finite signed measures on $(C_{n,q}, 2^{C_{n,q}})$ is a finite dimensional space and the set $\{\delta_x : x \in C_{n,q}\}$ is a basis for this space. Note that $\{\delta_x : x \in C_{n,q}\}$ is the set of all *point masses* on $C_{n,q}$, not vectors in the ambient space of $C_{n,q}$. Thus, to completely define the operator $T_{n,q}$, we need only define $T_{n,q}(\delta_x)$ for all $x \in C_{n,q}$. To this end, let $x \in C_{n,q}$. We define the function $F_{n,q} : C_{n,q} \to [q]^{\times n}$ as $F_{n,q}(x) = 1^{\times x_1} \times \cdots \times q^{\times x_q}$, where the exponents represent Cartesian powers. The definition of $F_{n,q}$ is a bit dense so we will do a simple example. Suppose $n = 6$, $q = 4$ and $x = [1, 0, 3, 2]^T$ then $F_{n,q}(x) = [1, 3, 3, 3, 4, 4]^T$. Intuitively, the $F_{n,q}$ operator undoes the totaling which transforms a collection of trials from a categorical distribution into a draw from a multinomial distribution; $F_{n,q}$ returns these trials in nondecreasing order. Let S_n be the symmetric group on n symbols. We define our linear operator as follows:

$$T_{n,q}(\delta_x) = \frac{1}{n!} \sum_{\sigma \in S_n} \delta_{\sigma(F_{n,q}(x))}, \tag{61}$$

where σ is permuting the entries of $F_{n,q}(x)$. This operator is similar to the projection operator onto the set of order n symmetric tensors [10]. The following lemma makes the crucial connection between the space of multinomial distributions and the probability measures of grouped samples.

LEMMA 7.1. *Let $Q_{n,p,q} \in \mathscr{M}(n, q)$, then*

$$T_{n,q}(Q_{n,p,q}) = V_n(\delta_{\sum_{i=1}^q p_i \delta_i}). \tag{62}$$

This lemma allows us to make some assertions about the identifiability of mixtures of multinomial distributions.

In the following, we will assume that all multinomial mixture models under consideration have only nonzero summands and distinct components. In the context of multinomial mixture models, a multinomial mixture model $\sum_{i=1}^m a_i Q_{n,p_i,q}$ is identifiable if it being equal to a different multinomial mixture model,

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q}, \tag{63}$$

with $s \leq m$ implies that $s = m$ and there exists some permutation σ such that $a_i = b_{\sigma(i)}$ and $Q_{n,p_i,q} = Q_{n,r_{\sigma(i)},q}$ for all i. The mixture model is determined if the previous statement holds without the restriction $s \leq m$.

Multinomial mixture models are identifiable if the number of components m and the number of trials in each component n satisfy $n \geq 2m - 1$.

COROLLARY 7.1. *Let $m \in \mathbb{N}_+$, $n \geq 2m - 1$, and fix $q \in \mathbb{N}_+$. Let $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}, Q_{n,r_1,q}, \ldots, Q_{n,r_s,q} \in \mathscr{M}(n, q)$ with $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ distinct,*

$Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ distinct, and $s \leq m$. If

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q} \tag{64}$$

with $a_i > 0$, $b_i > 0$ for all i and $\sum_{i=1}^{m} a_i = \sum_{j=1}^{s} b_j = 1$, then $s = m$ and there exists some permutation σ such that $a_i = b_{\sigma(i)}$ and $p_i = r_{\sigma(i)}$.

Alternatively, this corollary says that, given two different finite mixtures with components in $\mathscr{M}(n, q)$, one mixture with m components and the other with s components, if $n \geq 2m - 1$ and $n \geq 2s - 1$ then the mixtures induce different measures. Additionally, multinomial mixture models are determined if the number of components m and the number of trials in each component n satisfy $n \geq 2m$.

COROLLARY 7.2. *Let $n \geq 2m$ and fix $q \in \mathbb{N}$. Let $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ and $Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ be elements of $\mathscr{M}(n, q)$ with $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ distinct and $Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ distinct. If*

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q} \tag{65}$$

with $a_i > 0$, $b_i > 0$ for all i and $\sum_{i=1}^{m} a_i = \sum_{j=1}^{m} b_i = 1$, then $m = s$ and there exists some permutation σ such that $a_i = b_{\sigma(i)}$ and $p_i = r_{\sigma(i)}$.

Using the proof techniques employed in the proofs of these corollaries (see the Supplementary Material [25]), one could establish additional identifiability/determinedness style results for multinomial mixture models along the lines of Theorems 4.5, 4.6 and 4.7. Furthermore, it seems likely that one could use the algorithm described in the next section or from [2, 4, 22] to recover these components, using the transform $T_{n,q}$.

8. Algorithm. Here, we present an algorithm for the recovery of mixture components and proportions from data. The algorithm is quite general and can be applied to any measurable space. The Supplementary Material [25] contains a detailed description and analysis of the algorithm applied to categorical data, including a consistency proof.

Let $\sum_{i=1}^{m} w_i \delta_{\mu_i}$ be an arbitrary mixture of measures on some measurable space $(\Omega, \mathcal{F})$ which we are interested in recovering. Let $p_1, \ldots, p_m$ be square integrable densities with respect to a dominating measure ξ, with $\int_A p_i \, d\xi = \mu_i(A)$ for all $i \in [m]$ and $A \in \mathcal{F}$. A measure ξ and densities $p_1, \ldots, p_m$ satisfying these properties are guaranteed to exist as a consequence of Lemma 6.2.

We will consider the situation where we have $2m - 1$ samples per random group and have access to the tensors $\sum_{i=1}^{m} w_i p_i^{\otimes 2m-1}$ and $\sum_{i=1}^{m} w_i p_i^{\otimes 2m-2}$. In

a finite sample space, estimating these tensors is equivalent to estimating moment tensors of order $2m-1$ and $2m-2$. For measures over $\mathbb{R}^d$ dominated by the Lebesgue measure, one could estimate these tensors using a kernel density estimator in $\mathbb{R}^{d(2m-1)}$ and $\mathbb{R}^{d(2m-2)}$ using each sample group as a kernel center. We will also assume that $p_1, \ldots, p_m$ have distinct norms. Note that it is still possible to recover the mixture components if they do not have distinct norms. One way to do this is to choose ξ so that the norms are distinct. In the Supplementary Material [25], we describe a method which is guaranteed to do this when $(\Omega, \mathcal{F})$ is a finite sample space by choosing ξ randomly. We term this method "random dominating measure" in the experiments and supplement. Alternatively, if one is capable of choosing an element in $\operatorname{span}(\{p_1, \ldots, p_m\})$ in an appropriate random way, one could recover the mixture components using a variation of Jenrich's algorithm.

To describe the algorithm, we will need to make use of bounded linear operators on Hilbert spaces. Given a pair of Hilbert spaces H, H', we define $\mathcal{L}(H, H')$ as the space of *bounded linear operators* from H to H' and $\mathcal{L}(H) \triangleq \mathcal{L}(H, H)$. An operator, T, is in this space if there exists a nonnegative number C such that $\|Tx\|_{H'} \leq C\|x\|_H$ for all $x \in H$. The space of bounded linear operators is a Banach space when equipped with the norm

$$\|T\| \triangleq \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|}. \tag{66}$$

In addition, we will need to make use of tensor products of bounded linear operators. The following lemma is exactly Proposition 2.6.12 from [15].

LEMMA 8.1. *Let $H_1, \ldots, H_n$, $H_1', \ldots, H_n'$ be Hilbert spaces and let $U_i \in \mathcal{L}(H_i, H_i')$ for all $i \in [n]$. There exists a unique*

$$U \in \mathcal{L}\big(H_1 \otimes \cdots \otimes H_n, H_1' \otimes \cdots \otimes H_n'\big), \tag{67}$$

such that $U(h_1 \otimes \cdots \otimes h_n) = U_1(h_1) \otimes \cdots \otimes U_n(h_n)$ for all $h_1 \in H_1, \ldots, h_n \in H_n$.

DEFINITION 8.1. The operator constructed in Lemma 8.1 is called the *tensor product of* $U_1, \ldots, U_n$ and is denoted $U_1 \otimes \cdots \otimes U_n$.

Finally, we will need to employ Hilbert–Schmidt operators which are a subspace of the bounded linear operators.

DEFINITION 8.2. Let H, H' be Hilbert spaces and $T \in \mathcal{L}(H, H')$. T is called a *Hilbert–Schmidt operator* if $\sum_{x \in J} \|Tx\|^2 < \infty$ for an orthonormal basis $J \subset H$. We denote the set of Hilbert–Schmidt operators in $\mathcal{L}(H, H')$ by $\mathscr{HS}(H, H')$.

This definition does not depend on the choice of orthonormal basis: the sum $\sum_{x \in J} \|T(x)\|^2$ will always yield the same value regardless of the choice of orthonormal basis J.

Finally, we will also need to utilize the equivalence between tensor products and linear operators ([15], Proposition 2.6.9).

LEMMA 8.2. *Let H, H' be Hilbert spaces. There exists a unitary operator $U : H \otimes H' \to \mathscr{HS}(H, H')$ such that, for any simple tensor $h \otimes h' \in H \otimes H'$, $U(h \otimes h') = \langle h, \cdot \rangle h'$.*

Before we introduce the algorithm, we will discuss an important point regarding computational implementation and Lemmas 8.2 and 8.1. For the remainder of this paragraph, we will assume that Euclidean spaces are equipped with the standard inner product. Vectors in a space of tensor products of Euclidean spaces, for example, $\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_s}$ are easily represented on computers as elements of $\mathbb{R}^{d_1 \times \cdots \times d_s}$ [10]. Linear operators from some Euclidean tensor space to another can also be easily represented. Furthermore, the transformation in Lemma 8.2 and the construction of new operators from Lemma 8.1 can be implemented in computers by "unfolding" the tensors into matrices, applying common linear algebraic manipulations and "folding" them back into tensors. The inner workings of these manipulations are beyond the scope of this paper and we refer the reader to [14] for details. Practically speaking, this means the manipulations mentioned in Lemmas 8.2 and 8.1 are straightforward to implement with a bit of tensor programming know-how. Implementation may also be streamlined by using programming libraries that assist with these tensor manipulations such as the NumPy library for Python.

Because of the points mentioned in the previous paragraph, the following algorithm is readily implementable for estimating categorical distributions, where the measures can be represented as probability vectors on a Euclidean space. Similarly, we expect that these techniques could be extended to probability densities on Euclidean space using kernel density estimators with a kernel function with easily computable L^2 inner products (e.g., Gaussian kernels) although we suspect that implementation of such an algorithm may be significantly more involved.

To begin our description of the abstract algorithm, we will apply the transform from Lemma 8.2 to $\sum_{i=1}^m w_i p_i^{\otimes 2m-2}$ to get the operator

$$C = \sum_{i=1}^m w_i p_i^{\otimes m-1} \langle p_i^{\otimes m-1}, \cdot \rangle = \sum_{i=1}^m \sqrt{w_i} p_i^{\otimes m-1} \langle \sqrt{w_i} p_i^{\otimes m-1}, \cdot \rangle. \tag{68}$$

Here, C is a positive semidefinite (PSD) operator in $\mathcal{L}(L^2(\Omega, \mathcal{F}, \xi)^{\otimes m-1})$. Let $C^\dagger$ be the (Moore–Penrose) pseudoinverse of C and $W = \sqrt{C^\dagger}$. Now W is an operator that whitens $\sqrt{w_1} p_1^{\otimes m-1}, \ldots, \sqrt{w_m} p_m^{\otimes m-1}$. That is, $W\sqrt{w_1} p_1^{\otimes m-1}, \ldots, W\sqrt{w_m} p_m^{\otimes m-1}$ are orthonormal vectors. Using the operator construction from Lemma 8.1, we can construct $I \otimes W \otimes W$ where, for all simple tensors in $L^2(\Omega, \mathcal{F}, \xi)^{\otimes 2m-1}$ we have

$$\begin{aligned} &(I \otimes W \otimes W)(x_1 \otimes \cdots \otimes x_{2m-1}) \\ &\qquad = x_1 \otimes W(x_2 \otimes \cdots \otimes x_m) \otimes W(x_{m+1} \otimes \cdots \otimes x_{2m-1}). \end{aligned} \tag{69}$$

Applying $I \otimes W \otimes W$ to $\sum_{i=1}^{m} w_i p_i^{\otimes 2m-1}$ yields

$$A \triangleq \sum_{i=1}^{m} w_i p_i \otimes W(p_i^{\otimes m-1}) \otimes W(p_i^{\otimes m-1}) \tag{70}$$

$$= \sum_{i=1}^{m} p_i \otimes W(\sqrt{w_i} p_i^{\otimes m-1}) \otimes W(\sqrt{w_i} p_i^{\otimes m-1}). \tag{71}$$

From Lemma 8.2, we can transform the tensor A into the operator T,

$$T = \sum_{i=1}^{m} p_i \otimes W(\sqrt{w_i} p_i^{\otimes m-1}) \langle W(\sqrt{w_i} p_i^{\otimes m-1}), \cdot \rangle. \tag{72}$$

Because W is a whitening operator, the operator TT^H is

$$TT^H = \sum_{i=1}^{m} p_i \otimes W(\sqrt{w_i} p_i^{\otimes m-1}) \Bigg\langle W(\sqrt{w_i} p_i^{\otimes m-1}), \ldots \tag{73}$$

$$\sum_{j=1}^{m} W(\sqrt{w_j} p_j^{\otimes m-1}) \langle p_j \otimes W(\sqrt{w_j} p_j^{\otimes m-1}), \cdot \rangle \Bigg\rangle$$

$$= \sum_{i=1}^{m} p_i \otimes W(\sqrt{w_i} p_i^{\otimes m-1}) \langle p_i \otimes W(\sqrt{w_i} p_i^{\otimes m-1}), \cdot \rangle \tag{74}$$

which is a PSD operator. We set $S \triangleq TT^H$.

For $i \neq j$, it follows that $p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1} \perp p_j \otimes W\sqrt{w_j} p_j^{\otimes m-1}$. To see this,

$$\langle p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1}, p_j \otimes W\sqrt{w_j} p_j^{\otimes m-1} \rangle \tag{75}$$

$$= \langle p_i, p_j \rangle \langle W\sqrt{w_i} p_i^{\otimes m-1}, W\sqrt{w_j} p_j^{\otimes m-1} \rangle \tag{76}$$

$$= \langle p_i, p_j \rangle 0 \tag{77}$$

$$= 0. \tag{78}$$

Also note that

$$\| p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1} \|^2 = \langle p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1}, p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1} \rangle \tag{79}$$

$$= \langle p_i, p_i \rangle \langle W\sqrt{w_i} p_i^{\otimes m-1}, W\sqrt{w_i} p_i^{\otimes m-1} \rangle \tag{80}$$

$$= \| p_i \|^2. \tag{81}$$

If $p_1, \ldots, p_m$ have distinct norms, then it follows that

$$\sum_{i=1}^{m} p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1} \langle p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1}, \cdot \rangle \tag{82}$$

is the unique spectral decomposition of S since the vectors $p_1 \otimes W\sqrt{w_1}p_1^{\otimes m-1}$, $\ldots$, $p_m \otimes W\sqrt{w_m}p_m^{\otimes m-1}$ are orthogonal, have distinct norms, and thus distinct positive eigenvalues. Given an eigenvector of S, $p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}$, we need only view it as a linear operator $p_i\langle W\sqrt{w_i}p_i^{\otimes m-1}, \cdot\rangle$ and apply this operator to some vector z which is not orthogonal to $W\sqrt{w_i}p_i^{\otimes m-1}$, thus yielding p_i scaled by $\langle W\sqrt{w_i}p_i^{\otimes m-1}, z\rangle$. Where the norms of $p_1, \ldots, p_m$ not distinct, then there would not be a spectral gap between some of the eigenvalues in S, and a spectral decomposition of S may contain some eigenvectors that are not $p_1 \otimes W\sqrt{w_1}p_1^{\otimes m-1}, \ldots, p_m \otimes W\sqrt{w_m}p_m^{\otimes m-1}$, but are instead linear combinations of these vectors.

The following is a concise summary of the main points of the full algorithm:

1. Let $C = \sum_{i=1}^m w_i p_i^{\otimes m-1}\langle p_i^{\otimes m-1}, \cdot\rangle$ by transforming $\sum_{i=1}^m w_i p_i^{\otimes 2m-2}$.
2. Let $W = \sqrt{C^\dagger}$.
3. Let $A = I \otimes W \otimes W(\sum_{i=1}^m w_i p_i^{\otimes 2m-1})$. Note that

$$I \otimes W \otimes W\left(\sum_{i=1}^m w_i p_i^{\otimes 2m-1}\right) = \sum_{i=1}^m p_i \otimes W\big(\sqrt{w_i}p_i^{\otimes m-1}\big) \otimes W\big(\sqrt{w_i}p_i^{\otimes m-1}\big)$$

by direct evaluation and rearrangement of coefficients.

4. Let $T = \sum_{i=1}^m p_i \otimes W(\sqrt{w_i}p_i^{\otimes m-1})\langle W(\sqrt{w_i}p_i^{\otimes m-1}), \cdot\rangle$ by transforming A.
5. Performing spectral decomposition on TT^H gives us eigenvectors $\{p_i \otimes W(\sqrt{w_i}p_i^{\otimes m-1})\}_{i=1}^m$, up to scaling.
6. For all i, let $\tilde{p}_i = p_i\langle W(\sqrt{w_i}p_i^{\otimes m-1}), z\rangle$ by transforming the eigenvectors into linear operators and selecting z to be any vector such that the inner product does not evaluate to 0. Now $\tilde{p}_i$ is a scaled version of p_i.
7. Normalize $\tilde{p}_i$ to get p_i.

Once the mixture components $p_1, \ldots, p_m$ are recovered from the spectral decomposition we can calculate the mixture proportions. From these mixture components, we can construct the tensors $p_1^{\otimes 2m-2}, \ldots, p_m^{\otimes 2m-2}$. These tensors are linearly independent by Lemma 5.3. The tensor $\sum_{i=1}^m w_i p_i^{\otimes 2m-2}$ is known. By the linear independence of the components, there is exactly one solution for $a_1, \ldots, a_m$ in the equation

$$\sum_{i=1}^m w_i p_i^{\otimes 2m-2} = \sum_{j=1}^m a_j p_j^{\otimes 2m-2}, \tag{83}$$

so simply minimizing $\| \sum_{i=1}^m w_i p_i^{\otimes 2m-2} - \sum_{j=1}^m a_j p_j^{\otimes 2m-2}\|$ over $a_1, \ldots, a_m$ will give us the mixture proportions.

In the Supplementary Material [25], we study this algorithm applied to finite sample spaces in further detail. In the supplement, we demonstrate how to recover mixture components without the spectral gap assumption, how to construct the

estimator given data (which we evaluate experimentally in Section 9) and prove that it is consistent.

Taking inspiration from [2] and [23], we can suggest yet another algorithm. The previous papers demonstrate algorithms for recovering mixture components which are measures on finite sample spaces and $\mathbb{R}^d$, from random groups of size 3, provided the mixture components are linearly independent. Given a mixture of measures $\mathscr{P} = \sum_{i=1}^m w_i \delta_{\mu_i}$ with density functions $p_1, \ldots, p_m$, the tensors $p_1^{\otimes m-1}, \ldots, p_m^{\otimes m-1}$ are linearly independent. Thus, with $3m-3$ samples per random group, we can estimate the tensors $\sum_{i=1}^m w_i p_i^{\otimes 3m-3}$ and we can use the algorithms from the previous papers to recover $p_1^{\otimes m-1}, \ldots, p_m^{\otimes m-1}$ from which it is straightforward to recover $p_1, \ldots, p_m$.

9. Experiments. Here, we will present some experimental results of our algorithm applied to a simple synthetic dataset. The sample space for the experiments is $\Omega = \{0, 1, 2\}$. The mixture components of our dataset are μ_1, μ_2, μ_3 with μ_1 distributed according to a binomial distribution with $n = 2$ and $p = 0.2$, μ_2 is similar with $p = 0.8$ and $\mu_3 = \frac{1}{3}\mu_1 + \frac{2}{3}\mu_2$. The component weights are $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$. We chose these mixture components so that they are not particularly nice. Specifically, the mixture components are not linearly independent, and when considered as vectors in $\mathbb{R}^3$, μ_1 and μ_2 have the same norm. Our mixture of measures is $\mathscr{P} = \sum_{i=1}^3 w_i \delta_{\mu_i}$ and our samples come from either $V_5(\mathscr{P})$ or $V_6(\mathscr{P})$ depending on the algorithm used.

We construct our own performance measure which allows us to judge the performance of the estimated components jointly. Let $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3$ be the three estimates for the mixture components from some algorithm. We will view these estimates as vectors in $\mathbb{R}^3$. Our performance measure is $\min_{\sigma \in S_3} \frac{1}{3} \sum_{i=1}^3 \|\mu_i - \widehat{\mu}_{\sigma(i)}\|_{\ell^1(\mathbb{R}^3)}$. That is, we take the average of total variations of the best matching of the estimated mixture components to the true components.

9.1. *Proposed algorithms.* We include two different implementations of our proposed algorithm, one where we use what we call a "random dominating measure" and in the other we use what we call a "fixed dominating measure." In the following, we describe the two implementations and the rationale for presenting both of them.

In the description of our algorithm in Section 8, we make the assumption that the mixture components, when represented as square integrable densities over some dominating measure, have distinct L^2 norms. This is necessary to ensure that (82) admits a unique spectral decomposition. Because μ_1 and μ_2 have the same norm when considered as vectors in $\mathbb{R}^3$ this assumption does not hold for the experiments we present here. We use the aforementioned "random dominating measure" technique (details in the Supplementary Material [25]) which transforms the measure space so that the mixture components have distinct norms. To do this, we

choose a dominating measure randomly so that, with probability one, the mixture components have different norms when represented as densities over this measure space. We theoretically demonstrate that this technique works in the Supplementary Material [25]. In this paper, we present experimental evidence that this technique also works in practice.

The purpose of the random dominating measure is to create a spectral gap between the mixture components. Intuitively, it seems reasonable that if we choose the dominating measure "well" then we will end up with large spectral gaps without making any of the component norms so diminutive as to become unnoticeable. In the interest of exploring this idea, we tested different dominating measures until we found one that improved algorithmic performance significantly and include these experimental results as well. We found that the dominating measure ξ with $\xi(\{0\}) = 3^2$, $\xi(\{1\}) = 2^2$ and $\xi(\{2\}) = 1$ improved performance significantly and we refer to this as the "fixed dominating measure" implementation. These experimental results indicate the possibility for significant improvements to our algorithm by choosing the dominating measure intelligently. Additional specifics for these proposed implementations can be found in the Supplementary Material [25].

Both of these implementations were run on two experimental scenarios, one with 50,000 random groups and the other with 10,000,000 random groups, with all groups drawn from $V_5(\mathscr{P})$. We repeated each experiment 20 times and report relevant statistics.

9.2. *Competing algorithms*. As a baseline, we compare our algorithm against simply choosing 3 measures uniformly at random from the probabilistic simplex. The randomly selected components algorithm was repeated 1000 times. We also compare our algorithm to a modified version of the algorithm introduced in [2]. The algorithm in [2] is designed to work on random groups with three samples and a mixture of measures with linearly independent components. Because of this, we apply the algorithm in [2] to random groups from $V_6(\mathscr{P})$ rather than $V_5(\mathscr{P})$, with the adaptation described at the end of Section 8. This algorithm was also run on experimental scenarios with 50,000 and 10,000,000 random groups. Again, these experiments were repeated 20 times.

9.3. *Results*. The results are summarized in Table 1. Our algorithm demonstrates a clear improvement as the number of random groups increases. Our modification of the algorithm in [2] performs noticeably better than the other algorithms, likely owing to the fact that it has more information per group and/or the fact that it does not depend on the "random dominating measure" trick. Using the fixed dominating measure narrows this gap considerably, and it seems likely that this gap could be further improved with a better choice of dominating measure.

10. Discussion. In closing, we offer the following observations related to our results.

TABLE 1
Experimental results

Method	Performance
Random Dominating Measure, 50,000 samples	Mean: 0.1407, Variance: 0.0169
Fixed Dominating Measure, 50,000 samples	Mean: 0.0524, Variance: 0.0011
Anandkumar et al. [2], 50,000 samples	Mean: 0.0503, Variance: 0.0145
Random Dominating Measure, 10,000,000 samples	Mean: 0.0433, Variance: 0.0062
Fixed Dominating Measure, 10,000,000 samples	Mean: 0.0037, Variance: $4e{-}6$
Anandkumar et al. [2], 10,000,000 samples	Mean: 0.0026, Variance: $4e{-}6$
Randomly Selected Measures	Mean: 0.5323, Variance: 0.0203

10.1. *Potential statistical test and estimator.* The results on determinedness suggest the possibility of a goodness-of-fit test. Suppose we have grouped samples from some mixture of measures $\mathscr{P}' = \sum_{i=1}^{m'} w_i' \delta_{\mu_i'}$. Further suppose some null hypothesis

$$H_0 : \mathscr{P}' = \mathscr{P} \triangleq \sum_{i=1}^{m} w_i \delta_{\mu_i}. \tag{84}$$

Given data from $V_{2m}(\mathscr{P}')$, we may be able to reject the null hypothesis provided we have some way of estimating $M \triangleq \sum_{i=1}^{m} w_i \mu_i^{\times 2m}$ from the groups of samples. We will call such an estimator $\widehat{M}$. If $\widehat{M}$ does not converge to M, then we can reject the null hypothesis. The implementation and analysis of such an estimator would depend on the setting and is outside the scope of this paper

One interesting observation from the proof of Theorem 4.3 is that, if $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures, p_i is a pdf for μ_i for all i, and $n > m$, then the rank of $\sum_{i=1}^{m} a_i p_i^{\otimes n} \otimes p_i^{\otimes n}$ will be exactly m. This suggests a statistical estimator for the number of mixture components. The form of this tensor is amenable to spectral methods since it is a positive semidefinite tensor of order 2, which is akin to a positive semidefinite matrix. Embedding the data with the kernel mean mapping, using a universal kernel [19], seems like a promising approach to constructing such a test or estimator.

10.2. *Identifiability and the value* $2n - 1$. The value $2n - 1$ seems to carry some significance for identifiability beyond the setting we proposed. This value can also be found in results concerning metrics on trees [20], hidden Markov models [21] and frame theory, with applications to signal processing [6]. All of these results are related to identifiability of an object or the injectivity of an operator. We can offer no further insight as to why this value recurs, but it appears to be an algebraic phenomenon.

Acknowledgment. RV: Thanks to Marius Kloft for some interesting discussions.

SUPPLEMENTARY MATERIAL

Supplement to "An operator theoretic approach to nonparametric mixture models" (DOI: 10.1214/18-AOS1762SUPP; .pdf). Technical results and additional algorithmic details.

REFERENCES

[1] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

[2] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

[3] ANDERSON, J., BELKIN, M., GOYAL, N., RADEMACHER, L. and VOSS, J. (2014). The more, the merrier: The blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of the* 27*th Conference on Learning Theory* 1135–1164.

[4] ARORA, S., GE, R., KANNAN, R. and MOITRA, A. (2012). Computing a nonnegative matrix factorization—provably. In *STOC'*12*—Proceedings of the* 2012 *ACM Symposium on Theory of Computing* 145–161. ACM, New York. MR2961503

[5] ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models—going beyond SVD. In 2012 *IEEE* 53*rd Annual Symposium on Foundations of Computer Science—FOCS* 2012 1–10. IEEE Computer Soc., Los Alamitos, CA. MR3185945

[6] BALAN, R., CASAZZA, P. and EDIDIN, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.* **20** 345–356. MR2224902

[7] BHASKARA, A., CHARIKAR, M., MOITRA, A. and VIJAYARAGHAVAN, A. (2014). Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, *STOC'*14 594–603. ACM, New York.

[8] BLANCHARD, G. and SCOTT, C. (2014). Decontamination of mutually contaminated models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 1–9.

[9] BRUNI, C. and KOCH, G. (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.* **13** 1341–1357. MR0806230

[10] COMON, P., GOLUB, G., LIM, L.-H. and MOURRAIN, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.* **30** 1254–1279. MR2447451

[11] DONOHO, D. and STODDEN, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems* (S. Thrun, L. K. Saul and B. Schölkopf, eds.) **16** 1141–1148. MIT Press, Cambridge, MA.

[12] ELMORE, R. and WANG, S. (2003). Identifiability and estimation in finite mixture models with multinomial components. Technical Report 03-04, Dept. Statistics, Pennsylvania State Univ., State College, PA.

[13] FOLLAND, G. B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics* (*New York*). Wiley, New York. MR1681462

[14] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[15] KADISON, R. V. and RINGROSE, J. R. (1983). *Fundamentals of the Theory of Operator Algebras*. *Vol. I*: *Elementary Theory*. *Pure and Applied Mathematics* **100**. Academic Press, New York. MR0719020

[16] KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1876169

[17] KIM, B. S. (1984). Studies of multinomial mixture models. Ph.D. thesis, Univ. North Carolina, Chapel Hill.

[18] KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl*. **18** 95–138. MR0444690

[19] MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res*. **7** 2651–2667. MR2274454

[20] PACHTER, L. and SPEYER, D. (2004). Reconstructing trees from subtree weights. *Appl. Math. Lett*. **17** 615–621. MR2064171

[21] PAZ, A. (1971). *Introduction to Probabilistic Automata*. Academic Press, New York. MR0289222

[22] RABANI, Y., SCHULMAN, L. J. and SWAMY, C. (2014). Learning mixtures of arbitrary distributions over large discrete domains. In *ITCS'*14*—Proceedings of the* 2014 *Conference on Innovations in Theoretical Computer Science* 207–223. ACM, New York. MR3359477

[23] SONG, L., ANANDKUMAR, A., DAI, B. and XIE, B. (2014). Nonparametric estimation of multi-view latent variable models. In *Proceedings of the* 31*st International Conference on Machine Learning*, *ICML* 2014 640–648.

[24] TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat*. **34** 1265–1269. MR0155376

[25] VANDERMEULEN, R. A. and SCOTT, C. D. (2019). Supplement to "An operator theoretic approach to nonparametric mixture models." DOI:10.1214/18-AOS1762SUPP.

[26] YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Stat*. **39** 209–214. MR0224204

DEPARTMENT OF COMPUTER SCIENCE
FACHBEREICH INFORMATIK
TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN
POSTFACH 3049
67653 KAISERSLAUTERN
GERMANY
E-MAIL: vandermeulen@cs.uni-kl.de

ELECTRICAL AND COMPUTER
ENGINEERING, STATISTICS
UNIVERSITY OF MICHIGAN
EECS BUILDING
1301 BEAL AVENUE
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: clayscot@umich.edu