On The Identifiability of Mixture Models from Grouped Samples

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Finite mixture models are statistical models which appear in many problems in statistics and machine learning. In such models it is assumed that data are drawn from random probability measures, called mixture components, which are themselves drawn from a probability measure $\mathscr{P}$ over probability measures. When estimating mixture models, it is common to make assumptions on the mixture components, such as parametric assumptions. In this paper, we make no assumption 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 component. We show that any mixture of m probability measures can be uniquely identified provided there are $2m-1$ observations per group. Moreover we show that, for any m, there exists a mixture of m probability measures that cannot be uniquely identified when groups have $2m-2$ observations. Our results hold for any sample space with more than one element.

1 Introduction

A finite mixture model is a probability law based on a finite number of probability measures, $\mu_1, \ldots, \mu_m$, and a discrete distribution $w_1, \ldots, w_m$. A realization of a mixture model is ~~first~~ generated by first generating a component at random k, $1 \le k \le m$, and then drawing from μ_k. A mixture model can be associated with a probability measure on probability measures, which we denote $\mathscr{P}$ with $\mathscr{P}(\{\mu_i\}) = w_i$ for all i. Mixture models are used to model data throughout statistics and machine learning.

A primary theoretical question concerning mixture models is identifiability. A mixture model is said to be identifiable if no other mixture model (of equal or lesser complexity) explains the distribution of the data. Some previous work on identifiability considers the situation where the observations are drawn iid from the mixture model, and conditions on $\mu_1, \ldots, \mu_m$ are imposed, such as Gaussianity [7, 3]. In this work we make no assumptions on $\mu_1, \cdots, \mu_m$. Instead, we assume the observations are grouped, such that realizations from the same group are known to be iid from the same component. We call these groups of samples "random groups." We define a random group to be a random collection $\mathbf{X}_i$, where $\mathbf{X}_i = X_{i,1}, \ldots, X_{i,n} \overset{iid}{\sim} \nu_i$ and $\nu_i \overset{iid}{\sim} \mathscr{P}$. We call the setting where $\mathscr{P}$ is a probability measure over probability measures ($\mathscr{P}$ is not necessarily atomic) and $\mathbf{X}_i$ of the form described above, the *grouped sample setting*.

Consider the set of all mixtures of probability measures which yield the same distribution over the random groups as does $\mathscr{P}$. If some element of this set other than $\mathscr{P}$ has no more components than $\mathscr{P}$ then $\mathscr{P}$ is not identifiable. In other words, there is no way to differentiate $\mathscr{P}$ from another model of equal or lesser complexity. Fortunately, with a sufficient number of samples in each random group, $\mathscr{P}$ becomes the simplest model which describes the data. In this paper we show that, for any sample space, any mixture of probability measures with m components is identifiable when there are $2m - 1$ samples per random group. Furthermore we show that this bound cannot be improved, regardless of sample space.

not mentioned

1.1 Practical Implications of Results

Though a somewhat mathematically abstract object, probability measures over spaces of probability measures arise quite naturally in many statistical problems. Any application which use s mixture models, for example clustering, is utilizing a probability measure over probability measures. More recently there has been significant interest in the grouped sample setting. We will describe a few relevant machine learning problems here.

By far the weakest section. Perhaps move to a Discussion section at the end? Unrelated paragraphs.

How is the model of our paper related to transfer learning.

Transfer learning is the attempt to leverage several different but related training datasets to construct a classifier or regressor for another different but related testing dataset. In such a setting one may assume that there exists some probability measure over probability measures $\mathscr{P}$ which is generating unobserved random measures $(\nu_1, \nu_2, \ldots)$. From these random measures we have access to $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$. These groups of samples are our "different but related training datasets" with relatedness being induced through $\mathscr{P}$. Finally one may assume that the testing dataset is generated from $\tilde{\nu} \sim \mathscr{P}$ but with unobserved labels. This model has been used to construct effective transfer learning algorithms [4, 12].

I think it's important to acknowledge topic modeling.

Sometimes one would like to perform statistical techniques directly on a space of probability measures. Examples of this include detection of anomalous distributions [13] and distribution regression [14, 16]. Both of these problems assume a grouped sample setting.

In the grouped sample setting it is desirable to know how the number of samples in our random groups affects an algorithm's performance. For example: is having 10 samples per group enough to achieve satisfactory results for some application? In some settings group sample size may limited due to available data, but there are other considerations as well. For example in [2] large groups of samples are broken down to smaller groups of samples for algorithmic purposes. Furthermore, if we treat our random groups as one would typically treat a random vector, there may concerns that large random groups will cause issues because of the "curse of dimensionality." If this is the case it might make sense to break down the random groups into smaller random groups. What is lost by throwing away such information? This question is rather deep and will likely require substantial investigation before being solved in a general sort of way. Our work stands as a concrete step towards resolving such questions by answering the question, "if $\mathscr{P}$ is an atomic probability measure over probability measures with m atoms, what is the most samples per group we need to ensure that our data looks like it is generated uniquely from $\mathscr{P}$?" We also introduce a collection of new mathematical techniques for approaching such problems.

2 Related Work

The question of how many samples are necessary in each random group to uniquely identify a finite mixture of probability measures has come up sporadically over the past couple of decades. The application of Kruskal's theorem [11] has been used to concoct various identifiability results for random groups containing three samples. In [1] it was shown that any mixture of linearly independent measures over a discrete space or linearly independent probability distributions on $\mathbb{R}^d$ are identifiable from random groups containing three samples. The result most closely resembling our own is in [15]. This paper is primarily algorithmic; they introduce an algorithm for recovering m mixture components with $2m - 1$ samples per group in the discrete setting. They show that such an algorithm would not work for $2m - 2$ samples per group. Unfortunately their proof techniques are fairly difficult to penetrate and are inherently attached to the discrete setting. Our paper, on the other hand, is able to reach the same conclusion in a fully general setting using relatively simple arguments.

aren't there more refs? binomial case? Blischke Hettmansperger

find a "nicer" way to say

3 Problem Setup

Proofs of all the lemmas in the remainder of the paper can be found in the supplemental material. We will be treating this problem in as general of a setting as possible. For any measurable space we define δ_x as the Dirac measure at x. For ☺ a set, σ-algebra, or measure, we denote $☺^{\times a}$ to be the standard a-fold product associated with that object. For any natural number k we define $[k] \triangleq \mathbb{N} \bigcap [1, k]$. 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 $\mathcal{F} \neq \{\emptyset, \Omega\}$. We denote the space of probability measures over this space as $\mathcal{D}(\Omega, \mathcal{F})$, which we will shorten to $\mathcal{D}$. We will equip $\mathcal{D}$ with the σ-algebra $2^{\mathcal{D}}$ so that each Dirac measure over $\mathcal{D}$ is unique. Define $\Delta(\mathcal{D}) \triangleq \operatorname{span}(\delta_x : x \in \mathcal{D})$. This will be the ambient space where our mixtures of probability measures live. Let $\mathscr{P} = \sum_{i=1}^m \delta_{\mu_i} w_i$ be a probability measure in $\Delta(\mathcal{D})$. Let $\mu \sim \mathscr{P}$ and $X_1, \cdots, X_n \overset{iid}{\sim} \mu$. We will denote $\mathbf{X} = (X_1, \cdots, X_n)$. $\mathbf{X}$ is a random group from our mixture model.

We will now derive the probability law of $\mathbf{X}$. Let $A \in \Omega^{\times n}$, we have

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

The second equality follows from Lemma 3.10 in [10]. So the probability law of $\mathbf{X}$ is

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

We want to view the probability law of $\mathbf{X}$ as a function of $\mathscr{P}$ in a mathematically rigorous way, which requires a bit of technical buildup. Let $\mathcal{V}$ be a vector space. We will now construct a version of the integral for $\mathcal{V}$-valued functions over $\mathcal{D}$. Let $\mathscr{Q} \in \Delta(\mathcal{D})$. From the definition of $\Delta(\mathcal{D})$ it follows that $\mathscr{Q}$ admits the representation

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

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

this sentence doesn't fit in its paragraph. Can we just delete it?

Definition 1. We call $\mathscr{P}$ a *mixture of measures* if it is a probability measure in $\Delta(\mathcal{D})$. We will say that $\mathscr{P}$ has m *mixture components* if it has order m.

Lemma 1. *Let $\mathscr{Q} \in \Delta(\mathcal{D})$ ~~and~~ admit minimal representations $\mathscr{Q} = \sum_{i=1}^r \delta_{\mu_i} \alpha_i = \sum_{i=1}^r \delta_{\mu_i'} \alpha_i'$. There exists some permutation $\psi : [r] \to [r]$ such that $\mu_{\psi(i)} = \mu_i'$ and $\alpha_{\psi(i)} = \alpha_i'$ for all i.*

Henceforth when we define an element of $\Delta(\mathcal{D})$ with a summation we will assume that the summation is a minimal representation. 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 $\sum_{i=1}^m w_i = 1$. So any mixture of measures is a convex combination of Dirac measures at elements in $\mathcal{D}$.

For a function $f : \mathcal{D} \to \mathcal{V}$ define

$$\int f(\mu) d\mathscr{Q}(\mu) = \sum_{i=1}^r \alpha_i f(\mu_i),$$

where $\sum_{i=1}^r \delta_{\mu_i} \alpha_i$ is a minimal representation of $\mathscr{Q}$. This integral is well defined as a consequence of Lemma 1.

For a σ-algebra (Q, Σ) we define $\mathcal{M}(Q, \Sigma)$ as the space of all finite signed measures over that space. Let $\lambda_n : \mathcal{M}(\Omega, \mathcal{F}) \to \mathcal{M}(\Omega^{\times n}, \mathcal{F}^{\times n}); \mu \mapsto \mu^{\times n}$. We introduce the operator $V_n : \Delta(\mathcal{D}) \to \mathcal{M}(\Omega^{\times n}, \mathcal{F}^{\times n})$

$$V_n(\mathscr{Q}) = \int \lambda_n(\mu) d\mathscr{Q}(\mu) = \int \mu^{\times n} d\mathscr{Q}(\mu).$$

For a minimal representation $\mathscr{Q} = \sum_{i=1}^r \delta_{\mu_i} \alpha_i$, we have

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

From this definition we have that $V_n(\mathscr{P})$ is simply the law of $\mathbf{X}$ which we derived earlier. Two mixtures of measures are different if they admit a different measure over $\mathcal{D}$.

what does this mean?

Definition 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 2 is the central object 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. In topic modelling $\mathbf{X}$ would be the samples from a single document and in transfer learning it would be one of the several collections of training samples. If $\mathscr{P}$ is not n-identifiable then we know that there exists a mixture of measures which is no more complex (in terms of number of mixture components) than $\mathscr{P}$ which is not discernible from $\mathscr{P}$ given the data. 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 $\mathbf{X}$.

not mentioned

why is n-identifiability interesting in the context of transfer learning?

4 Results

Our primary result gives us a bound on the n-identifiability of ~~all~~ mixtures of measures with m or fewer components. We also show that this bound is tight.

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

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

~~Unsurprisingly, if a mixture of measures is n-identifiable then it is q-identifiable for all $q > n$. Likewise if a mixture of measures is not n-identifiable then it is not q-identifiable for $q < n$. Thus~~ identifiability is, in some sense, monotonic.

The following lemmas convey the unsurprising fact that

Lemma 2. *If a mixture of measures is n-identifiable then it is q-identifiable for all $q > n$.*

Lemma 3. *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 all minimal mixtures of measures with m or fewer components.

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. To our knowledge there does not exist a rigorous construction of the tensor product Hilbert space which is both succinct and intuitive. Because of this 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 through treatment of tensor products of Hilbert spaces can be found in [9].

↖o

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 treat the simple tensors as being the basis for some inner product space H_0, with the inner product of simple tensors satisfying

$$\langle h_1 \otimes h_1', h_2 \otimes h_2' \rangle = \langle h_1, h_2 \rangle \langle h_1', h_2' \rangle .$$

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 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 $L^2(\mathbb{R}^2)$ which maps the simple tensors to separable functions, $f \otimes f' \mapsto f(\cdot)f'(\cdot)$. 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

isomorphisms are onto, but not all elements of $L^2(\mathbb{R}^2)$ are separable.

us

$$\begin{aligned}\int\int (f(x)f'(y))\,(g(x)g'(y))dxdy &= \int f(x)g(x)dx \int f'(y)g'(y)dy \\ &= \langle f, g\rangle \langle f', g'\rangle \\ &= \langle f\otimes f', g\otimes g'\rangle .\end{aligned}$$

Beyond tensor product we will need to define tensor power. To begin we will first show that tensor products are, in some sense, associative. Let H_1, H_2, H_3 be Hilbert spaces. Proposition 2.6.5 in [9] states that there is a unique unitary operator, $U : (H_1 \otimes H_2) \otimes H_3 \to H_1 \otimes (H_2 \otimes H_3)$, which satisfies the following for all $h_1 \in H_1, h_2 \in H_2, h_3 \in H_3$,

$$U\left((h_1 \otimes h_2) \otimes h_3\right) = h_1 \otimes (h_2 \otimes h_3)\,.$$

This implies that for any collection of Hilbert spaces, $H_1, \cdots, H_n$, 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 [9] 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 $H^{\otimes n} = \underbrace{H \otimes H \otimes \cdots \otimes H}_{\text{n times}}$ and for $h \in H$, $h^{\otimes n} = \underbrace{h \otimes h \otimes \cdots \otimes h}_{\text{n times}}$.

5.2 Some Results for Tensor Product Spaces

Here we will state technical results which 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. Let $(\Psi, \mathcal{G}, \mu)$ be a σ-finite measure space. We have the following lemma which connects the L^2 space of products of measures to the tensor products of the L^2 space for each measure. This following lemma allows us to treat products of functions in L^2 as tensor products of functions.

Lemma 4. *There exists a unitary transform $U : L^2(\Psi, \mathcal{G}, \mu)^{\otimes n} \to L^2(\Psi^{\times n}, \mathcal{G}^{\times n}, \mu^{\times n})$ such that, for all $f_1, \cdots, f_n \in L^2(\Psi, \mathcal{G}, \mu)$, $U(f_1 \otimes \cdots \otimes f_n) = f_1(\cdot) \cdots f_n(\cdot)$.*

Unitary ⇒ invertible, but once again this map is not onto.

The following lemma used in the proof of Lemma 4 as well as the proof of Theorem 2.

Lemma 5. *Let $H_1, \cdots, H_n, H'_1, \cdots, H'_n$ be a collection of Hilbert spaces and $U_1, \cdots, U_n$ 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 $h_1 \in H_1, \cdots, h_n \in H_n$.*

A statement of the next lemma for $\mathbb{R}^d$ can be found in [6]. We present our own proof for the Hilbert space setting in the supplemental material.

Lemma 6. *Let $n > 1$ and let $h_1, \cdots, h_n$ be elements of a Hilbert space H such that no elements are zero and no pairs of elements are collinear. Then $h_1^{\otimes n-1}, \cdots h_n^{\otimes n-1}$ are linearly independent.* in $H^{\otimes n-1}$ (?)

6 Proofs of Theorems

With the tools developed in the previous sections we can now prove our theorems. First we introduce one additional piece of notation. For a function p on a domain $\mathcal{X}$ we define $p^{\times k}$ as simply the product of the function k times on the domain $\mathcal{X}^{\times k}$, $\underbrace{p(\cdot) \cdots p(\cdot)}_{\text{k times}}$. For a measure the notation continues to denote the standard product measure.

Finally will need the following technical lemma to connect the product of Radon-Nikodym derivatives to product measures.

Lemma 7. *Let $(\Psi, \mathcal{G})$ be a measurable space, η and γ a pair of bounded measures on that space, and f a nonnegative function in $L^1(\gamma)$ such that, for all $A \in \mathcal{G}$, $\eta(A) = \int_A f d\gamma$. Then for all n, for all $B \in \mathcal{G}^{\times n}$ we have*

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

Proof of Theorem 1. We will proceed by contradiction. Suppose there exist two different mixtures of measures $\mathscr{P} = \sum_{i=1}^{l} \delta_{\mu_i} a_i \neq \mathscr{P}' = \sum_{j=1}^{m} \delta_{\nu_j} b_j$, such that

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

and $l \leq m$. From our assumption on representation we know $\mu_i \neq \mu_j$ for all $i \neq j$ and similarly for $\nu_1, \cdots, \nu_m$. We will also assume that $\mu_i \neq \nu_j$ for all i, j. Were this not true we could simply subtract the smaller of the common terms from both sides of (2) and normalize to yield another pair of distinct mixtures of measures with fewer components and no shared terms, $\mathscr{Q}, \mathscr{Q}'$. Letting $\mathscr{Q}$ have m' components and $\mathscr{Q}'$ have l' with $m' \geq l'$ applying Lemma 3 would give us $V_{2m'-1}(\mathscr{Q}) = V_{2m'-1}(\mathscr{Q}')$ and we could proceed as usual.

Let $\xi = \sum_{i=1}^{l} \mu_i + \sum_{j=1}^{m} \nu_j$. Clearly ξ dominates μ_i and ν_j for all i, j so we can define Radon-Nikodym derivatives $p_i = \frac{d\mu_i}{d\xi}$, $q_j = \frac{d\nu_j}{d\xi}$ which are in $L^1(\Omega, \mathcal{F}, \xi)$. We can assert that these derivatives are everywhere nonnegative without issue. Clearly no two of these derivatives are equal. If one of the derivatives were a scalar multiple of another, for example $p_1 = \alpha p_2$ for some $\alpha \neq 1$, it would imply

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

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

Lemma 7 tells us that, for any $R \in \mathcal{F}^{\times 2m-1}$ we have

$$\begin{aligned} \int_R \sum_{i=1}^{l} a_i p_i^{\times 2m-1} d\xi^{\times 2m-1} &= \sum_{i=1}^{l} a_i \mu_i^{\times 2m-1}(R) \\ &= \sum_{j=1}^{m} b_j \nu_j^{\times 2m-1}(R) \\ &= \int_R \sum_{j=1}^{m} b_j q_j^{\times 2m-1} d\xi^{\times 2m-1}. \end{aligned}$$

Therefore

$$\sum_{i=1}^{l} a_i p_i^{\times 2m-1} = \sum_{j=1}^{m} b_j q_j^{\times 2m-1} \tag{3}$$

$\xi^{\times 2m-1}$-almost everywhere (Proposition 2.23 in [8]). We will now show for all i, j that $p_i \in L^2(\Omega, \mathcal{F}, \xi)$ and $q_j \in L^2(\Omega, \mathcal{F}, \xi)$. We will argue this for p_1 which will clearly generalize to the other elements. First we will show that $p_1 \leq 1$ ξ-almost everywhere. Suppose this were not true and that there exists $A \in \mathcal{F}$ with $\xi(A) > 0$ and $p_1(A) > 1$. Now we would have

$$\mu_1(A) = \int_A p_1 d\xi > \int_A 1 d\xi = \xi(A) = \sum_{i=1}^{l} \mu_i(A) + \sum_{j=1}^{m} \nu_j(A) \geq \mu_1(A)$$

a contradiction. Evaluating directly we get

$$\begin{aligned} \int p_1(\omega)^2 d\xi(\omega) &\leq \int 1 d\xi(\omega) \\ &= \xi(\Omega) \\ &= l + m, \end{aligned}$$

so $p_1 \in L^2(\Omega, \mathcal{F}, \xi)$. Applying the U^{-1} operator from Lemma 4 to (3) yields

$$\sum_{i=1}^{l} a_i p_1^{\otimes 2m-1} = \sum_{j=1}^{m} b_j q_j^{\otimes 2m-1}.$$

Since $l+m \leq 2m$ Lemma 6 states that $p_1^{\otimes 2m-1}, \cdots, p_l^{\otimes 2m-1}, q_1^{\otimes 2m-1}, \cdots, q_m^{\otimes 2m-1}$ are all linearly independent and thus $a_i = 0$ and $b_j = 0$ for all i, j, a contradiction. □

Proof of Theorem 2. To prove this theorem we will construct a pair of different mixture of measures, $\mathscr{P} \neq \mathscr{P}'$ which both contain m components and satisfy $V_{2m-2}\left(\mathscr{P}\right) = V_{2m-2}\left(\mathscr{P}'\right)$.

From our definition of $(\Omega, \mathcal{F})$ we know there exists $F \in \mathcal{F}$ such that F, F^C are nonempty. Let $f \in F$ and $f' \in F^C$. It follows that $\delta_f \neq \delta_{f'}$ are different probability measures on $(\Omega, \mathcal{F})$. Because δ_f and $\delta_{f'}$ are dominated by $\xi = \delta_f + \delta_{f'}$ we know that there exists a pair of measurable functions p, p' such that, for all A, $\delta_f\left(A\right) = \int_A p d\xi$ and $\delta_{f'}\left(A\right) = \int_A p' d\xi$. We can assert that p and p' are nonnegative without issue.

From the same argument we used in the proof of Theorem 1 we know $p, p' \in L^2\left(\Omega, \mathcal{F}, \xi\right)$. Let H_2 be the Hilbert space generated from the span of p, p'. Let $(\varepsilon_i)_{i=1}^{2m}$ be $2m$ distinct elements of $[0,1]$ and let $(p_i)_{i=1}^{2m}$ be elements of $L^1(\Omega, \mathcal{F}, \xi)$ with $p_i = \varepsilon_i p + (1 - \varepsilon_i) p'$. Clearly p_i is a pdf over ξ for all i and there are no pairs in this collection which are collinear. Let H_2 be the Hilbert space generated from the span of p and p'. Since H_2 is isomorphic to $\mathbb{R}^2$ there exists a unitary operator $U : H_2 \to \mathbb{R}^2$. From Lemma 5 there exists a unitary operator $U_{2m-2} : H_2^{\otimes 2m-2} \to \mathbb{R}^{2^{\otimes 2m-2}}$ with $U_{2m-2}\left(h_1 \otimes \cdots \otimes h_{2m-2}\right) = U(h_1) \otimes \cdots \otimes U(h_{2m-2})$. Because U is unitary the set $U_{2m-2}\left(\operatorname{span}\left(\left\{h^{\otimes 2m-2} : h \in H_2\right\}\right)\right)$ maps exactly to the set span $\left(x^{\otimes 2m-2} : x \in \mathbb{R}^2\right)$. An order r tensor, $A_{i_1,\ldots,i_r}$, is *symmetric* if $A_{\psi(i_1),\ldots,\psi(i_r)} = A_{i_1,\ldots,i_r}$for any $i_1, \cdots, i_r$ and permutation ψ. A consequence of Lemma 4.2 in [6] is that span $\left(\left\{x^{\otimes 2m-2} : x \in \mathbb{R}^2\right\}\right) \subset S^{2m-2}(\mathbb{C}^2)$ is exactly the space of all symmetric order $2m-2$ tensors over $\mathbb{C}^2$.

From Proposition 3.4 in [6] it follows that the dimension of $S^{2m-2}\left(\mathbb{C}^2\right)$ is $\begin{pmatrix} 2+2m-2-1 \\ 2m-2 \end{pmatrix} = 2m-1$. From this we get that $\dim\left(\operatorname{span}\left(\left\{h^{\otimes 2m-2} : h \in H_2\right\}\right)\right) \leq 2m-1$.

The bound on the dimension of span $\left(\left\{h^{\otimes 2m-2} : h \in H_2\right\}\right)$ implies that $\left(p_i^{\otimes 2m-2}\right)_{i=1}^{2m}$ are linearly dependent. Conversely Lemma 6 implies that removing a single vector from $\left(p_i^{\otimes 2m-2}\right)_{i=1}^{2m}$ yields a set of vectors which are linearly independent. It follows that there exists $(\alpha_i)_{i=1}^{2m}$ with $\alpha_i \neq 0$ for all i and

$$\sum_{i=1}^{2m} \alpha_i p_i^{\otimes 2m-2} = 0.$$

Without loss of generality we will assume that $\alpha_i < 0$ for $i \in [k]$ with $k \leq m$. From this we have

$$\sum_{i=1}^{k} -\alpha_i p_i^{\otimes 2m-2} = \sum_{j=k+1}^{2m} \alpha_j p_j^{\otimes 2m-2}. \tag{4}$$

From Lemma 4 we have

$$\sum_{i=1}^{k} -\alpha_i p_i^{\times 2m-2} = \sum_{j=k+1}^{2m} \alpha_j p_j^{\times 2m-2}$$

and thus

$$\begin{aligned} \int \sum_{i=1}^{k} -\alpha_i p_i^{\times 2m-2} d\xi^{\times 2m-2} &= \int \sum_{j=k+1}^{2m} \alpha_j p_j^{\times 2m-2} d\xi^{\times 2m-2} \\ \Rightarrow \sum_{i=1}^{k} -\alpha_i &= \sum_{j=k+1}^{2m} \alpha_j. \end{aligned}$$

Let $r = \sum_{i=1}^{k} -\alpha_i$. We know $r > 0$ so dividing both sides of (4) by r gives us

$$\sum_{i=1}^{k} -\frac{\alpha_i}{r} p_i^{\otimes 2m-2} = \sum_{j=k+1}^{2m} \frac{\alpha_j}{r} p_j^{\otimes 2m-2}$$

and the left and the right side are convex combinations. Let $(\beta_i)_{i=1}^{2m}$ positive numbers with $\beta_i = \frac{-\alpha_i}{r}$ for $i \in \{1, \cdots, k\}$ and $\beta_j = \frac{\alpha_j}{r}$ for $j \in \{k+1, \cdots, 2m\}$. This gives us

$$\sum_{i=1}^{k} \beta_i p_i^{\otimes 2m-2} = \sum_{j=k+1}^{2m} \beta_j p_j^{\otimes 2m-2}.$$

It follows that

$$\sum_{i=1}^{k} \beta_i p_i^{\otimes m-1} \otimes p_i^{\otimes m-1} = \sum_{j=k+1}^{2m} \beta_j p_j^{\otimes m-1} \otimes p_i^{\otimes m-1}.$$

We will now show that $k = m$. Suppose $k < m$. Then $p_1^{\otimes m-1}, \cdots, p_{k+1}^{\otimes m-1}$ are linearly independent. From this we know that there exists z such that $z \perp p_i^{\otimes m-1}$ for $i \in [k]$ but z is not orthogonal to $p_{k+1}^{\otimes m-1}$. Using this vector we have

$$\begin{aligned}\left\langle \sum_{i=1}^{k} \beta_i p_i^{\otimes 2m-1}, z \otimes z \right\rangle &= \sum_{i=1}^{k} \beta_i \left\langle z, p_i^{\otimes m-1} \right\rangle \left\langle z, p_i^{\otimes m-1} \right\rangle \\ &= 0\end{aligned}$$

but

$$\begin{aligned}\left\langle \sum_{i=k+1}^{2m} \beta_i p_i^{\otimes m-1} \otimes p_i^{\otimes m-1}, z \otimes z \right\rangle &= \sum_{j=k+1}^{2m} \beta_i \left\langle p_i^{\otimes m-1}, z \right\rangle \left\langle p_i^{\otimes m-1}, z \right\rangle \\ &> 0\end{aligned}$$

and thus $k = m$. Now we have

$$\sum_{i=1}^{m} \beta_i p_i^{\otimes 2m-2} = \sum_{j=m+1}^{2m} \beta_j p_j^{\otimes 2m-2}.$$

Applying Lemma 4 we get that

$$\sum_{i=1}^{m} \beta_i p_i^{\times 2m-2} = \sum_{j=m+1}^{2m} \beta_j p_j^{\times 2m-2}.$$

From Lemma 7 we have,

$$\sum_{i=1}^{m} \beta_i \left(\varepsilon_i \delta_f + (1-\varepsilon_i)\, \delta_{f'}\right)^{\times 2m-2} = \sum_{j=m+1}^{2m} \beta_j \left(\varepsilon_j \delta_f + (1-\varepsilon_j)\, \delta_{f'}\right)^{\times 2m-2}.$$

Setting $\mu_i = (\varepsilon_i \delta_f + (1-\varepsilon_i)\, \delta_{f'})$ yields

$$\sum_{i=1}^{m} \beta_i \mu_i^{\times 2m-2} = \sum_{j=m+1}^{2m} \beta_j \mu_j^{\times 2m-2}.$$

Thus setting $\mathscr{P} = \sum_{i=1}^{m} \beta_i \delta_{\mu_i}$ and $\mathscr{P}' = \sum_{j=m+1}^{2m} \beta_j \delta_{\mu_j}$ gives us $V_{2m-2}(\mathscr{P}) = V_{2m-2}(\mathscr{P}')$ and $\mathscr{P} \neq \mathscr{P}'$ by construction. □

Point out that identifiability is a ~~[illegible]~~ prerequisite to the study of algorithms.

7 Potential Algorithm and Conclusion

When considering the tensor product structure used in our proofs, it seems likely that, with $2m$ samples per group, one could use spectral methods to recover the mixture components from an empirical estimate of

$$\sum_{i=1}^{2m} a_i p_i^{\otimes m} \otimes p_i^{\otimes m} \cong \sum_{i=1}^{2m} a_i p_i^{\otimes m} \left\langle p_i^{\otimes m}, \cdot \right\rangle.$$

One could represent the elements p_i using a kernel density estimator or a reproducing kernel Hilbert space embedding of the data with universal kernels [5]. We have no further suggestions on how to construct such an algorithm, but such an approach seems promising.

In this paper we have proven a fundamental bound on the identifiability of mixture models in a nonparametric setting. Any mixture with m components is identifiable with groups of samples containing $2m-1$ samples from the same latent probability measure. We show that this bound is tight by constructing a mixture of m probability measures which is not identifiable with groups of samples containing $2m-2$. These results hold for any mixture over any domain with at least two elements.

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] 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.

[3] 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.

[4] 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.

[5] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *in Advances in Neural Information Processing Systems*, pages 406–414, 2010.

[6] 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.

[7] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *J. Mach. Learn. Res.*, 8:203–226, May 2007.

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

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

[10] Olav Kallenberg. *Foundations of modern probability*. Probability and its applications. Springer, New York, Berlin,, Paris, 2002. Sur la 4e de couv. : This new edition contains four new chapters as well as numerous improvements throughout the text.

[11] 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.

[12] Andreas Maurer, Massi Pontil, and Bernardino Romera-paredes. Sparse coding for multitask and transfer learning. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 343–351. JMLR Workshop and Conference Proceedings, May 2013.

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

[14] 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.

[15] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. *ArXiv e-prints*, 2013.

[16] Z. Szabo, B. Sriperumbudur, B. Poczos, and A. Gretton. Learning Theory for Distribution Regression. *ArXiv e-prints*, November 2014.

AISTATS 2015