
Supplement to Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations

Anonymous Author(s)

Affiliation

Address

email

1	Contents	
2	S1 Additional Experimental Details	2
3	S1.1 Out-of-sample Results	2
4	S1.2 Preprocessing of Twitter Dataset	2
5	S2 Optimization Details	3
6	S2.1 Full Optimization Problem	3
7	S2.2 Coreset Approach	4
8	S2.3 Algorithm	5
9	S2.4 Spectral Initialization	5
10	S3 Proof and General Form of Theorem 1	7
11	S3.1 Proof of Theorem 1: Groups of Size Two	7
12	S3.2 Theorem 1: Arbitrary Group Size	9
13	S3.2.1 Preliminaries	9
14	S3.2.2 Proof of Theorem 1: Arbitrary Group Size	10
15	S4 Theorem 2	11
16	S4.1 Intermediate Results	11
17	S4.2 Proof of Theorem 2: Groups of Size Two	12
18	S4.3 Proof of Theorem 2: Arbitrary Group Size	13
19	S5 Background on the Grouped Sample Setting and Proof of Theorem 3	15
20	S5.1 Identifiability in the Grouped Sample Setting	15
21	S5.2 Notation	16
22	S5.3 Full Theorem Statement and Proof	16
23	S6 General Version of Corollary 1	21

24 S1 Additional Experimental Details

25 In this section we provide details of the Twitter experiment and out-of-sample results for experiments
26 on synthetic datasets.

27 S1.1 Out-of-sample Results

28 Here we provide out-of-sample results for the synthetic experiments shown in the main paper. These
29 are shown in Table 1. We make the realistic assumption that pair information is not available for
30 out-of-sample data. We generate a test dataset 20% the size of the training set according to the
31 distribution of the training data.

Table 1: Out-of-sample ARI and standard deviation over 10 runs on synthetic datasets.

Dataset	NDIGO	CSC	NPMIX	CGMM	MVLVM
Overlapping moons	0.705 ± 0.091	0.405 ± 0.283	0.131 ± 0.075	0.002 ± 0.011	0.593 ± 0.229
Olympic Rings	0.607 ± 0.058	0.387 ± 0.033	0.367 ± 0.063	0.127 ± 0.012	0.290 ± 0.053
Half-disks	0.221 ± 0.036	0.215 ± 0.038	0.102 ± 0.095	0.185 ± 0.076	0.127 ± 0.062

32 S1.2 Preprocessing of Twitter Dataset

33 Twitter dataset is publicly available through FiveThirtyEight.¹ The data consist of tweets, from a
34 variety Russian-troll twitter accounts, tweeted between 2015 and 2018. We considered all tweets from
35 2016, a total of 878, 878. We pre-processed the tweets by removing stop words, punctuation, and
36 hyperlinks, followed by a lemmatization step and the removal of any words that were not contained
37 in the vocabulary of the six billion token GloVe word vectors [9]. For completeness, we mention
38 that lemmatization is a common pre-processing step in natural language processing that removes
39 inflectional differences from words by mapping each inflection to a common base form called the
40 lemma. For example, lemmatization will map each of the words dog, dogs, dog’s, dogs’, and doggy
41 to the word dog. For each tweet, we paired the constituent words uniformly at random without
42 replacement, resulting in 1, 691, 081 pairs of words where the words of a given pair come from the
43 same tweet. No other information was retained. We emphasize that a given word from a given tweet
44 will *not* be assigned to more than one pair. However, if a given word appears in multiple tweets
45 (which may not all be about the same topic), it will show up in multiple pairs.

46 For the embedding step we performed PCA on the pre-trained GloVe 50-dimensional embeddings to
47 obtain 10-dimensional vectors which were used to encode the paired words. The kernel centers were
48 obtained by running mini-batch k -means with $R = 200$ on a uniform random sample of 10, 000 of
49 the 1, 691, 081 word pairs, from which the matrix G was also calculated. We then trained on the full
50 1, 691, 081 word pairs, which are utilized in mini-batches through the matrix $C^{(t)}$.

51 Algorithms for competing methods, as described by their respective authors, could not scale to this
52 experiment. Therefore, we compare to recent methods designed for continuous topic modeling
53 of short texts: LF-DMM [8], and GPU-DMM[6] as implemented by Qiang et al. [10].² LF-
54 DMM and GPU-DMM were trained on the same preprocessed data as NDIGO, where each of the
55 1, 691, 081 word pairs is considered a unique document. Each of these methods were run with
56 default hyperparameters, as described in the documentation for GPU-DMM and LF-DMM.² After
57 training, UCI topic coherence [7] (which measures pointwise mutual information) was used to
58 evaluate performance. UCI topic coherence uses a reference dataset to estimate word co-occurrence
59 probabilities, which is more robust in the short text setting as very common words in a given topic
60 may never be observed to co-occur. A recent Wikipedia article dump was used for the reference
61 dataset, and is provided with our code.

¹<https://github.com/fivethirtyeight/russian-troll-tweets>

²<https://github.com/qiang2100/STTM>

62 S2 Optimization Details

63 In this section we provide details of our algorithm, including the form of the objective function and
 64 initialization, for both the full problem and the coreset approach.

65 Recall the expression for the ETISE

$$\hat{J}(w, \alpha) \triangleq \int q_{w, \alpha}^2(x, x') dx dx' - 2 \sum_{m=1}^M \sum_{r=1}^n \sum_{r'=1}^2 \sum_{s=1}^n \sum_{s'=1}^2 w_m \alpha_{mr, r'} \alpha_{ms, s'} \hat{h}(r, r', s, s'),$$

66 where

$$\begin{aligned} \hat{h}(r, r', s, s') &\triangleq \begin{cases} \hat{h}_{\text{LTO}}(r, r', s, s'), & r \neq s \\ \hat{h}_{\text{LOO}}(r, r', s'), & r = s \end{cases} \\ \hat{h}_{\text{LOO}}(r, r', r'') &\triangleq \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} k_{\sigma}(x_{i,1}, x_{r, r'}) k_{\sigma}(x_{i,2}, x_{r, r''}) \\ \hat{h}_{\text{LTO}}(r, r', s, s') &\triangleq \frac{1}{n-2} \sum_{i \in [n] \setminus \{r, s\}} k_{\sigma}(x_{i,1}, x_{r, r'}) k_{\sigma}(x_{i,2}, x_{s, s'}). \end{aligned}$$

67 For ease of computation, we will rewrite $\hat{J}(w, \alpha)$ in terms of matrix operations. In what follows, we
 68 assume that $\tilde{k}_{\sigma}(z_r, z_u) := \int k_{\sigma}(x, z_r) k_{\sigma}(x, z_u) dx$ has a closed-form expression or can otherwise be
 69 computed efficiently. Some examples [5] are give in Table 2.

Table 2: Some popular kernel functions and their associated \tilde{k} . Here $\|\cdot\|_2$ is the Euclidean norm.

Kernel	$k_{\sigma}(x, x')$	$\tilde{k}_{\sigma}(x, x')$
Gaussian	$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\ x-x'\ _2^2}{2\sigma^2}\right)$	$k_{\sqrt{2}\sigma}(x, x')$
Cauchy	$\left(\frac{1}{\sqrt{\pi}\sigma}\right)^d \left(\frac{\Gamma((1+d)/2)}{\Gamma(1/2)}\right) \left(\frac{\sigma^2 + \ x-x'\ _2^2}{\sigma^2}\right)^{-\frac{1+d}{2}}$	$k_{2\sigma}(x, x')$
Laplacian	$\frac{c_d}{\sigma^d} \exp\left(-\frac{\ x-x'\ _1}{\sigma}\right)$	$\frac{1}{(4\sigma)^d} \prod_{l=1}^d \left(\frac{\sigma + x_l - x'_l }{\sigma}\right) \exp\left(-\frac{\ x-x'\ _1}{\sigma}\right)$

70 S2.1 Full Optimization Problem

71 We begin by examining the first term of $\hat{J}(w, \alpha)$

$$\begin{aligned} \int q_{w, \alpha}(x, x')^2 dx dx' &= \int \left(\sum_m w_m \sum_r \sum_{r'} \alpha_{m, r, r'} k_{\sigma}(x, x_{r, r'}) \sum_s \sum_{s'} \alpha_{m, s, s'} k_{\sigma}(x, x_{s, s'}) \right) \\ &\quad \times \left(\sum_j w_j \sum_u \sum_{u'} \alpha_{j, u, u'} k_{\sigma}(x, x_{u, u'}) \sum_v \sum_{v'} \alpha_{j, v, v'} k_{\sigma}(x, x_{v, v'}) \right) dx dx' \\ &= \sum_{m, j} w_m w_j \sum_{r, r', u, u'} \sum_{s, s', v, v'} \alpha_{m, r, r'} \alpha_{m, s, s'} \alpha_{j, u, u'} \alpha_{j, v, v'} \\ &\quad \times \int k_{\sigma}(x, x_{r, r'}) k_{\sigma}(x, x_{u, u'}) dx \int k_{\sigma}(x', x_{s, s'}) k_{\sigma}(x', x_{v, v'}) dx' \\ &= \sum_{m, j} w_m w_j \sum_{r, r', u, u'} \alpha_{m, r, r'} \alpha_{j, u, u'} \tilde{k}_{\sigma}(x_{r, r'}, x_{u, u'}) \sum_{s, s', v, v'} \alpha_{m, s, s'} \alpha_{j, v, v'} \tilde{k}_{\sigma}(x_{s, s'}, x_{v, v'}) \\ &= \sum_{m, j} w_m w_j \left(\alpha'_m G a_j \right)^2, \end{aligned}$$

72 where \times in the first line is scalar multiplication and G is the kernel matrix of the data and is given by

$$G = \begin{bmatrix} \tilde{k}_\sigma(x_{1,1}, x_{1,1}) & \tilde{k}_\sigma(x_{1,1}, x_{1,2}) & \cdots & \cdots & \tilde{k}_\sigma(x_{1,1}, x_{n,1}) & \tilde{k}_\sigma(x_{1,1}, x_{n,2}) \\ \tilde{k}_\sigma(x_{1,2}, x_{1,1}) & \tilde{k}_\sigma(x_{1,2}, x_{1,2}) & \cdots & \cdots & \tilde{k}_\sigma(x_{1,2}, x_{n,1}) & \tilde{k}_\sigma(x_{1,2}, x_{n,2}) \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \tilde{k}_\sigma(x_{n,1}, x_{1,1}) & \tilde{k}_\sigma(x_{n,1}, x_{1,2}) & \cdots & \cdots & \tilde{k}_\sigma(x_{n,1}, x_{n,1}) & \tilde{k}_\sigma(x_{n,1}, x_{n,2}) \\ \tilde{k}_\sigma(x_{n,2}, x_{1,1}) & \tilde{k}_\sigma(x_{n,2}, x_{1,2}) & \cdots & \cdots & \tilde{k}_\sigma(x_{n,2}, x_{n,1}) & \tilde{k}_\sigma(x_{n,2}, x_{n,2}) \end{bmatrix}.$$

73 Examining the second term of the ETISE yields

$$\begin{aligned} \mathbb{E}_q [q_{w,\alpha}] &\approx \sum_{m=1}^M \sum_{r=1}^n \sum_{r'=1}^2 \sum_{s=1}^n \sum_{s'=1}^2 w_m \alpha_{m,r,r'} \alpha_{m,s,s'} \hat{h}(r, r', s, s') \\ &= \begin{cases} \frac{1}{n-2} \sum_m \sum_{r,r'} \sum_{s,s'} w_m \alpha_{m,r,r'} \alpha_{m,s,s'} \sum_{i \in [n] \setminus \{r,s\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{s,s'}), & r \neq s \\ \frac{1}{n-1} \sum_m \sum_{r,r'} \sum_{s,s'} w_m \alpha_{m,r,r'} \alpha_{m,s,s'} \sum_{i \in [n] \setminus \{r\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{r,s'}), & r = s \end{cases} \\ &= \sum_{m=1}^M w_m (\alpha'_m C \alpha_m), \end{aligned}$$

74 where C is given by

$$C = \begin{bmatrix} \hat{h}(1, 1, 1, 1) & \hat{h}(1, 1, 1, 2) & \cdots & \cdots & \hat{h}(1, 1, n, 1) & \hat{h}(1, 1, n, 2) \\ \hat{h}(1, 2, 1, 1) & \hat{h}(1, 2, 1, 2) & \cdots & \cdots & \hat{h}(1, 2, n, 1) & \hat{h}(1, 2, n, 2) \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \hat{h}(n, 1, 1, 1) & \hat{h}(n, 1, 1, 2) & \cdots & \cdots & \hat{h}(n, 1, n, 1) & \hat{h}(n, 1, n, 2) \\ \hat{h}(n, 2, 1, 1) & \hat{h}(n, 2, 1, 2) & \cdots & \cdots & \hat{h}(n, 2, n, 1) & \hat{h}(n, 2, n, 2) \end{bmatrix}.$$

75 The diagonal blocks of size two of the matrix C use the leave one out estimator, while the other
76 entries use the leave two out estimator.

77 S2.2 Coreset Approach

78 In the full problem the matrices $G, C \in \mathbb{R}^{2n \times 2n}$ grow linearly with the data. This can make the
79 proposed optimization problem (6) costly to solve, as the complexity of gradient calculations are
80 quadratic in the dimensions of G, C . Additionally, the complexity of evaluating out-of-sample data is
81 quadratic in n for general KDEs. The motivation of the coreset approach is to reduce this complexity.

82 KDEs traditionally center kernels at the location of each observation, i.e., $k_\sigma(\cdot, x_{i,i'})$, where we
83 call $x_{i,i'}$ the kernel center. Rather than constraining the wKDE to have kernels centered at the
84 observations, we can formulate the optimization problem with R kernel centers $z_r \in \mathbb{R}^d$ for some
85 suitably chosen z_r . Additionally, choosing $R \ll n$ will substantially reduce the complexity of
86 gradient calculations and out-of-sample evaluation. The collection of kernel centers z_r will be our
87 coreset. We don't provide guarantees for the optimality of any particular coreset. The coreset could
88 potentially be chosen as the cluster centers output by some clustering algorithm, some suitable subset
89 of the data, or perhaps via some more principled scheme. In all of our experiments, we chose the
90 coreset to be cluster centers output by mini-batch k -means, where the number of clusters was chosen
91 to be $R > M$.

92 For the coreset approach, the ETISE has the same form but the matrices G and C have the form

$$G = \begin{bmatrix} \tilde{k}_\sigma(z_1, z_1) & \cdots & \cdots & \tilde{k}_\sigma(z_1, z_R) \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{k}_\sigma(z_R, z_1) & \cdots & \cdots & \tilde{k}_\sigma(z_R, z_R) \end{bmatrix},$$

$$C = \frac{1}{n} \sum_{i=1}^n C_i = \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} k_\sigma(x_{i,1}, z_1)k_\sigma(x_{i,2}, z_1) & \cdots & \cdots & k_\sigma(x_{i,1}, z_1)k_\sigma(x_{i,2}, z_R) \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ k_\sigma(x_{i,1}, z_R)k_\sigma(x_{i,2}, z_1) & \cdots & \cdots & k_\sigma(x_{i,1}, z_R)k_\sigma(x_{i,2}, z_R) \end{bmatrix}$$

93 This is derived in the same way as the full problem, replacing the kernel centers $x_{r,r'}, x_{s,s'}$ with the
94 coreset z_r , and using the "leave-none-out" estimator in place of the LOO/LTO estimator \hat{h} .

95 S2.3 Algorithm

96 Though the problem (7) is nonconvex, we observe that a properly initialized alternating pro-
97 jected stochastic gradient descent (APSGD) procedure produces good solutions in practice. Pseu-
98 docode for the APSGD algorithm for solving (7) is given in Algorithm 1. We mention that
99 the projections Π_Δ are onto the probability simplex, a decaying step size $\eta^{(t)}$ is used, and
100 stochasticity is introduced via the matrix $C^{(t)}$, which is a mini-batch version of C defined by
101 $C_{a,b}^{(t)} = \frac{1}{|\Omega^{(t)} \setminus \{a,b\}|} \sum_{i \in |\Omega^{(t)} \setminus \{a,b\}|} k_\sigma(X_{i,1}, X_{\lfloor \frac{a}{2} \rfloor, a \bmod 2}) k_\sigma(X_{i,2}, X_{\lfloor \frac{b}{2} \rfloor, b \bmod 2})$, where $\Omega^{(t)}$ is the
102 index set corresponding to the t^{th} mini-batch.

Algorithm 1 Alternating Projected SGD

```

1: init:  $\alpha^{(0)}, w^{(0)}, \eta^{(0)}$ 
2: procedure APSGD( $\alpha^{(0)}, w^{(0)}, \eta^{(0)}$ )
3:   Form  $G$ 
4:   for  $t = 1, 2, \dots$  do
5:     Take a minibatch of paired observations indexed by  $\Omega^{(t)}$ 
6:     Form  $C^{(t)}$  from the minibatch according to the definition of  $C$ 
7:      $w^{(t)} = \Pi_\Delta(w^{(t-1)} - \eta^{(t)} \nabla_w \hat{J}(w^{(t-1)}, \alpha^{(t-1)}))$ 
8:     for  $j = 1, \dots, M$  do
9:        $\alpha_j^{(t)} = \Pi_\Delta(\alpha_j^{(t-1)} - \eta^{(t)} \nabla_{\alpha_j} \hat{J}(w^{(t)}, \alpha_j^{(t-1)}))$ 

```

103 S2.4 Spectral Initialization

We adopt a spectral initialization scheme. First, the initialization is presented for the full problem, then we adapt it to the coreset approach. The idea here is, given some estimator of q , let's say \tilde{q} , to find a low rank approximation of \tilde{q}

$$\tilde{q}(x, y) \approx \sum_{i=1}^M \lambda_i \psi_i(x) \psi_i(y)$$

104 and then to use λ_i and ψ_i as starting points for our mixture weights and components. We do this by
105 using the full grouped sample data as an estimate of q which we transform into a linear operator and
106 decompose using a functional eigenvector decomposition.

107 We begin with a standard KDE applied to our full samples using a product kernel:

$$f_\sigma(y, y') = \frac{1}{2n} \sum_{i=1}^n k_\sigma(y, x_{i,1})k_\sigma(y', x_{i,2}) + k_\sigma(y, x_{i,2})k_\sigma(y', x_{i,1}).$$

108 Note that we include centers at both $(x_{i,1}, x_{i,2})$ and $(x_{i,2}, x_{i,1})$ so our KDE is symmetric in y, y' .

By Lemmas 5.1 and 8.2 of Vandermeulen and Scott [12], f_σ can be viewed as an element of a tensor product space $L^2(\mathbb{R}^d) \otimes L^2(\mathbb{R}^d)$ as follows

$$f_\sigma = \frac{1}{2n} \sum_{i=1}^n k_\sigma(\cdot, x_{i,1}) \otimes k_\sigma(\cdot, x_{i,2}) + k_\sigma(\cdot, x_{i,2}) \otimes k_\sigma(\cdot, x_{i,1}).$$

By the Lemmas referenced above, there is a unitary transformation on the KDE f_σ such that it can be viewed as a linear operator $T : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ given by

$$T(g) := \sum_{i=1}^n k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), g(\cdot) \rangle_{L^2} + k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), g(\cdot) \rangle_{L^2}, \quad \forall g \in L^2(\mathbb{R}^d)$$

which is a symmetric operator since it includes both $k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), g(\cdot) \rangle_{L^2}$ and $k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), g(\cdot) \rangle_{L^2}$ terms. We have removed the $1/(2n)$ coefficient since it will not affect the spectral decomposition. For any $g \in L^2$ the quantity $\langle k_\sigma(\cdot, x_{i,i'}), g(\cdot) \rangle_{L^2}$ will be a finite scalar, so $T(g)$ will be a linear combination of the $k_\sigma(\cdot, x_{i,i'})$. Therefore, eigenvectors of the above linear operator will have the form $g(\cdot) = \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'})$ since T applied to any vector must lie in the span of $k_\sigma(\cdot, x_{j,j'})$. Evaluating T on vectors of this form (not necessarily an eigenvector) will yield

$$\begin{aligned} T(g) &:= \sum_{i=1}^n \left\{ k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'}) \rangle_{L^2} \right. \\ &\quad \left. + k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'}) \rangle_{L^2} \right\} \\ &= \sum_i \zeta_{i,1} k_\sigma(\cdot, x_{i,1}) + \zeta_{i,2} k_\sigma(\cdot, x_{i,2}) \end{aligned}$$

where $\zeta_{i,1} = \sum_{j,j'} \beta_{j,j'} \tilde{k}_\sigma(x_{i,2}, x_{j,j'})$, and $\zeta_{i,2} = \sum_{j,j'} \beta_{j,j'} \tilde{k}_\sigma(x_{i,1}, x_{j,j'})$ and $\tilde{k}_\sigma(y, y') := \langle k_\sigma(\cdot, y), k_\sigma(\cdot, y') \rangle_{L^2}$.

Define the ordering of the elements of β and ζ by

$$\begin{aligned} \beta &= [\beta_{1,1}, \beta_{1,2}, \beta_{2,1}, \beta_{2,2}, \dots, \beta_{n,1}, \beta_{n,2}]', \\ \zeta &= [\zeta_{1,1}, \zeta_{1,2}, \zeta_{2,1}, \zeta_{2,2}, \dots, \zeta_{n,1}, \zeta_{n,2}]'. \end{aligned}$$

Then we have

$$\zeta = \bar{G}\beta,$$

where

$$\bar{G} = \begin{bmatrix} \tilde{k}_\sigma(x_{1,2}, x_{1,1}) & \tilde{k}_\sigma(x_{1,2}, x_{1,2}) & \tilde{k}_\sigma(x_{1,2}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{1,2}, x_{n,1}) & \tilde{k}_\sigma(x_{1,2}, x_{n,2}) \\ \tilde{k}_\sigma(x_{1,1}, x_{1,1}) & \tilde{k}_\sigma(x_{1,1}, x_{1,2}) & \tilde{k}_\sigma(x_{1,1}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{1,1}, x_{n,1}) & \tilde{k}_\sigma(x_{1,1}, x_{n,2}) \\ \tilde{k}_\sigma(x_{2,2}, x_{1,1}) & \tilde{k}_\sigma(x_{2,2}, x_{1,2}) & \tilde{k}_\sigma(x_{2,2}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{2,2}, x_{n,1}) & \tilde{k}_\sigma(x_{2,2}, x_{n,2}) \\ \tilde{k}_\sigma(x_{2,1}, x_{1,1}) & \tilde{k}_\sigma(x_{2,1}, x_{1,2}) & \tilde{k}_\sigma(x_{2,1}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{2,1}, x_{n,1}) & \tilde{k}_\sigma(x_{2,1}, x_{n,2}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{k}_\sigma(x_{n,2}, x_{1,1}) & \tilde{k}_\sigma(x_{n,2}, x_{1,2}) & \tilde{k}_\sigma(x_{n,2}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{n,2}, x_{n,1}) & \tilde{k}_\sigma(x_{n,2}, x_{n,2}) \\ \tilde{k}_\sigma(x_{n,1}, x_{1,1}) & \tilde{k}_\sigma(x_{n,1}, x_{1,2}) & \tilde{k}_\sigma(x_{n,1}, x_{2,1}) & \cdots & \tilde{k}_\sigma(x_{n,1}, x_{n,1}) & \tilde{k}_\sigma(x_{n,1}, x_{n,2}) \end{bmatrix}$$

is a row permutation of G defined for the full problem, obtained by exchanging rows corresponding to the first and second elements of each paired observation. The takeaway is that the coefficients of the eigenvectors of T are given by the right eigenvectors of \bar{G} . In particular, we will take the right eigenvectors of \bar{G} corresponding to the M largest real eigenvalues with the intuition that they will capture the dominant modes of T . Note that these eigenvectors will contain real-valued entries by the spectral theorem since T is symmetric. We call these eigenvectors the first, second, and so on. It should be noted that \bar{G} is not a symmetric matrix and so eigenvectors should be found according to, for example, the power iteration or orthogonal iteration. In general the eigenvectors of \bar{G} will have negative entries and not sum to one, so we project these eigenvectors onto the probability simplex to obtain the non-negative weights for initialization where we take α_1 to be the projection of the first eigenvector of \bar{G} , α_2 to be the second, and so on. For the initial w_i , We take w_1 to be the first eigenvalue of \bar{G} , w_2 to be the second and so on, projecting the resulting $w = [w_1, w_2, \dots, w_M]'$ onto the probability simplex.

136 **Coreset Approach Initialization** Initialization for the coreset approach is similar. Since we assume
 137 no relationship between z_i , we don't have the same notion of using paired kernel centers even though
 138 our data is still paired. However, we can still write the KDE using a product kernel over the coreset as

$$f_\sigma(y, y') = \frac{1}{R} \sum_{i=1}^R k_\sigma(y, z_i) k_\sigma(y', z_i).$$

This KDE is symmetric since the same kernel center is used in each term of the product kernel. Again appealing to Lemmas 5.1 and 8.2 of Vandermeulen and Scott [12], f_σ can be viewed as an element of a tensor product space $L^2(\mathbb{R}^d) \otimes L^2(\mathbb{R}^d)$ as follows

$$f_\sigma = \frac{1}{R} \sum_{i=1}^R k_\sigma(\cdot, z_i) \otimes k_\sigma(\cdot, z_i).$$

139 By the Lemmas referenced above, there is a unitary transformation on the KDE f_σ such that it can be
 140 viewed as a linear operator $T : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ given by

$$T(g) = \sum_{i=1}^R k_\sigma(\cdot, z_i) \langle k_\sigma(\cdot, z_i), g(\cdot) \rangle_{L^2}, \quad \forall g \in L^2.$$

141 By the same argument as the full problem, the eigenvectors of the above operator have the form
 142 $g(\cdot) = \sum_j \beta_j k_\sigma(\cdot, z_j)$. Applying T to a vector of this form (not necessarily an eigenvector), we have

$$\begin{aligned} T\left(\sum_j \beta_j k_\sigma(\cdot, z_j)\right) &= \sum_{i=1}^R k_\sigma(\cdot, z_i) \langle k_\sigma(\cdot, z_i), \sum_j \beta_j k_\sigma(\cdot, z_j) \rangle_{L^2} \\ &= \sum_{i=1}^R \zeta_i k_\sigma(\cdot, z_i), \end{aligned}$$

143 where $\zeta_i = \sum_j \beta_j \langle k_\sigma(\cdot, z_i), k_\sigma(\cdot, z_j) \rangle$. In this setting we have the standard ordering, $\beta = [\beta_1, \beta_2, \dots, \beta_R]'$ and
 144 $\zeta = [\zeta_1, \zeta_2, \dots, \zeta_R]'$. In matrix form, the relationship between β and ζ is given by

$$\zeta = G\beta,$$

145 where G is as previously defined for the coreset approach. No row permutation is needed in this
 146 setting as both centers of our product kernel are the same. From this point, the initialization scheme
 147 is essentially the same as for the full problem, but using the eigenvectors and eigenvalues of G . One
 148 key difference is that G is a symmetric matrix, so the eigenvalues and eigenvectors of G can be found
 149 using any standard solver.

150 S3 Proof and General Form of Theorem 1

151 Theorem 1 was presented in the main paper for groups of size $N = 2$. Here, we provide the proof for
 152 groups of size two, as well as the proof for the more general case of arbitrary group size $N > 2$. One
 153 tool we will use is Hoeffding's inequality for independent bounded random variables, which we state
 154 here for completeness.

155 **Theorem.** *Hoeffding's Inequality: Let V_1, V_2, \dots, V_n be independent bounded random variables*
 156 *such that $a_i \leq V_i \leq b_i$ with probability one. If $S_n = \sum_{i=1}^n V_i$, then for all $t > 0$*

$$P\left\{\left|S_n - \mathbb{E}\{S_n\}\right| \geq t\right\} \leq 2 \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

157 S3.1 Proof of Theorem 1: Groups of Size Two

158 We restate Theorem 1 for convenience.

159 **Theorem 1.** *Let $\epsilon > 0$ and set $\delta = 8(n^2 - n) \exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\} + 8n \exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}$. With*
 160 *probability at least $1 - \delta$ the following holds:*

$$\|q - q_{\hat{w}, \hat{\alpha}}\|_2^2 \leq \inf_{w \in \Delta^M, \alpha \in \Delta_{2n}^{2n}} \|q - q_{w, \alpha}\|_2^2 + \epsilon.$$

161 *Proof.* Our goal is to bound $|J(w, a) - \hat{J}(w, a)|$ uniformly over $w \in \Delta^M$, $\alpha \in \Delta_M^{2n}$. Recall the
 162 following definitions

$$\begin{aligned} h(r, r', s, s') &:= \int k_\sigma(x, x_{r,r'}) k_\sigma(x', x_{s,s'}) q(x, x') dx dx' \\ \hat{h}(r, r', s, s') &:= \begin{cases} \hat{h}_{\text{LTO}}(r, r', s, s'), & r \neq s \\ \hat{h}_{\text{LOO}}(r, r', s'), & r = s \end{cases} \\ \hat{h}_{\text{LOO}}(r, r', s') &:= \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{r,r'}) \\ \hat{h}_{\text{LTO}}(r, r', s, s') &:= \frac{1}{n-2} \sum_{i \in [n] \setminus \{r,s\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{s,s'}). \end{aligned}$$

163 The use of the leave one out (LOO) and leave two out (LTO) estimators above is to ensure indepen-
 164 dence so that we will be able to apply Hoeffding's inequality. We have

$$\begin{aligned} &P_q \left\{ \sup_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} |J(w, \alpha) - \hat{J}(w, \alpha)| > \frac{\epsilon}{2} \right\} \\ &= P_q \left\{ \sup_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} \left| \sum_{m=1}^M w_m \sum_{r=1}^n \sum_{s=1}^n \sum_{r'=1}^2 \sum_{s'=1}^2 \alpha_{m,r,r'} \alpha_{m,s,s'} h(r, r', s, s') \right. \right. \\ &\quad \left. \left. - \sum_{m=1}^M w_m \sum_{r=1}^n \sum_{s=1}^n \sum_{r'=1}^2 \sum_{s'=1}^2 \alpha_{m,r,r'} \alpha_{m,s,s'} \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\} \\ &\leq P_q \left\{ \sup_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} \sum_m \sum_{r,s} \sum_{r',s'} w_m \alpha_{m,r,r'} \alpha_{m,s,s'} \left| h(r, r', s, s') - \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\} \\ &\leq P_q \left\{ \max_{r,s,r',s'} \left| h(r, r', s, s') - \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\} \\ &\leq \sum_{r,s} \sum_{r',s'} P_q \left\{ \left| h(r, r', s, s') - \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\}. \end{aligned}$$

165 The second step above is due to the triangle inequality, and the penultimate step is due to simplex
 166 constraints on w, α . Let $k_i(r, r', s, s') := k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{s,s'})$. Noting that $h(r, r', s, s') =$
 167 $\mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\}$,

$$\begin{aligned} &P_q \left\{ \left| h(r, r', s, s') - \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\} \\ &= \begin{cases} P_q \left\{ \left| \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - \frac{1}{n-2} \sum_{i \in [n] \setminus \{r,s\}} k_i(r, r', s, s') \right| > \frac{\epsilon}{4} \right\}, & r \neq s \\ P_q \left\{ \left| \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} k_i(r, r', s, s') \right| > \frac{\epsilon}{4} \right\}, & r = s \end{cases} \\ &= \begin{cases} P_q \left\{ \left| \frac{1}{n-2} \sum_{i \in [n] \setminus \{r,s\}} \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - k_i(r, r', s, s') \right| > \frac{\epsilon}{4} \right\}, & r \neq s \\ P_q \left\{ \left| \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - k_i(r, r', s, s') \right| > \frac{\epsilon}{4} \right\}, & r = s \end{cases} \\ &= \begin{cases} P_q \left\{ \left| \sum_{i \in [n] \setminus \{r,s\}} \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - k_i(r, r', s, s') \right| > \frac{(n-2)\epsilon}{4} \right\}, & r \neq s \\ P_q \left\{ \left| \sum_{i \in [n] \setminus \{r\}} \mathbb{E}_{(x_{i,1}, x_{i,2}) \sim q} \{k_i(r, r', s, s')\} - k_i(r, r', s, s') \right| > \frac{(n-1)\epsilon}{4} \right\}, & r = s \end{cases} \end{aligned}$$

168 The terms $k_i(r, r', s, s')$ are independent random variables due to use of the LOO/LTO estimator.
 169 By assumption, $0 \leq k_i(r, r', s, s') \leq C_k^2 \sigma^{-2d}$ so the k_i are bounded for fixed $\sigma > 0$. We apply

170 Hoeffding's inequality

$$P_q \left\{ \left| h(r, r', s, s') - \hat{h}(r, r', s, s') \right| > \frac{\epsilon}{4} \right\} \leq \begin{cases} 2 \exp\left\{-\frac{2(n-2)^2 \epsilon^2}{16(n-2)C_k^4 \sigma^{-4d}}\right\}, & r \neq s \\ 2 \exp\left\{-\frac{2(n-1)^2 \epsilon^2}{16(n-1)C_k^4 \sigma^{-4d}}\right\}, & r = s \end{cases} \\ \leq \begin{cases} 2 \exp\left\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\right\}, & r \neq s \\ 2 \exp\left\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\right\}, & r = s \end{cases}. \quad (\text{S1})$$

171 Substituting backward we obtain the desired upper bound

$$P_q \left\{ \sup_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} |J(w, \alpha) - \hat{J}(w, \alpha)| > \frac{\epsilon}{2} \right\} \leq \sum_{r,s} \sum_{r',s'} \begin{cases} 2 \exp\left\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\right\}, & r \neq s \\ 2 \exp\left\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\right\}, & r = s \end{cases} \\ = 8(n^2 - n) \exp\left\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\right\} + 8n \exp\left\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\right\}.$$

172 Letting $\delta = 8(n^2 - n) \exp\left\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\right\} + 8n \exp\left\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\right\}$, we have

$$J(w, \alpha) - \frac{\epsilon}{2} \leq \hat{J}(w, \alpha) \leq J(w, \alpha) + \frac{\epsilon}{2} \quad \forall w, \alpha \quad (\text{S2})$$

173 with probability at least $1 - \delta$. Thus, with probability at least $1 - \delta$, for any $w \in \Delta^M, \alpha \in \Delta_M^{2n}$

$$\begin{aligned} J(\hat{w}, \hat{\alpha}) &\leq \hat{J}(\hat{w}, \hat{\alpha}) + \frac{\epsilon}{2} \\ &\leq \hat{J}(w, \alpha) + \frac{\epsilon}{2} \\ &\leq J(w, \alpha) + \epsilon, \end{aligned}$$

174 where $\hat{w}, \hat{\alpha}$ are defined in (6). Then with probability at least $1 - \delta$

$$\hat{J}(\hat{w}, \hat{\alpha}) \leq \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} J(w, \alpha) + \epsilon. \quad (\text{S3})$$

175 Combining (S3) with the definition of the ISE shows, with probability at least $1 - \delta$,

$$\|q - q_{\hat{w}, \hat{\alpha}}\|_2^2 \leq \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} \|q - q_{w, \alpha}\|_2^2 + \epsilon.$$

176 □

177 S3.2 Theorem 1: Arbitrary Group Size

178 S3.2.1 Preliminaries

179 Before beginning the proof, we start by redefining $q, q_{w, \alpha}, J$, and \hat{J} for arbitrary group size. Once
180 this is done, the proof will follow the same basic steps as the proof for groups of size two.

181 Suppose we change the problem setup only in the size of the grouped observations. Consider
182 grouped observations of size N . Consider a set of n grouped observations $\mathbf{x}_1, \dots, \mathbf{x}_n$ with $\mathbf{x}_i =$
183 $(x_{i,1}, \dots, x_{i,N}) \in \mathbb{R}_N^d := \underbrace{\mathbb{R}^d \times \dots \times \mathbb{R}^d}_N$ drawn i.i.d. from

$$q(y_1, y_2, \dots, y_N) = \sum_{m=1}^M w_m^* p_m^*(y_1) p_m^*(y_2) \dots p_m^*(y_N), \quad y_1, y_2, \dots, y_N \in \mathbb{R}^d. \quad (\text{S4})$$

184 Similar to the paired observation setting, a wKDE in this setting will have the form

$$p(y; \theta) = \sum_{r=1}^n \sum_{r'=1}^N \theta_{r,r'} k_\sigma(y, x_{r,r'}).$$

185 We may write the corresponding estimator of q

$$q_{w,\alpha}(y_1, y_2, \dots, y_N) = \sum_{m=1}^M w_m p(y_1; \alpha_m) p(y_2; \alpha_m) \dots p(y_N; \alpha_m)$$

186 where $\alpha_m = [\alpha_{m,1,1} \dots \alpha_{m,1,N} \dots \alpha_{m,n,1} \dots \alpha_{m,n,N}]' \in \Delta^{Nn}$ for $m = 1, \dots, M$, with
 187 $\alpha_{m,r,r'}$ corresponding to the weight of the kernel centered at $x_{r,r'}$ in the estimate of the m^{th} mixture
 188 component.

189 In what follows we use $\sum_{r,r'} := \sum_{r_1,r'_1} \dots \sum_{r_N,r'_N}$ to ease notation. Similar to the paired sample
 190 case, we define

$$J(w, \alpha) := \int q_{w,\alpha}^2(y_1, \dots, y_N) dy_1 \dots dy_N - 2 \sum_{m,r,r'} \left(\prod_{i \in [N]} w_m \alpha_{m,r_i,r'_i} \right) h(r_1, r'_1, \dots, r_N, r'_N)$$

$$\hat{J}(w, \alpha) := \int q_{w,\alpha}^2(y_1, \dots, y_N) dy_1 \dots dy_N - 2 \sum_{m,r,r'} \left(\prod_{i \in [N]} w_m \alpha_{m,r_i,r'_i} \right) \hat{h}(r_1, r'_1, \dots, r_N, r'_N),$$

191 where

$$h(r_1, r'_1, \dots, r_N, r'_N) := \int k_\sigma(y_1, x_{r_1,r'_1}) \dots k_\sigma(y_N, x_{r_N,r'_N}) q(y_1, \dots, y_N) dy_1 \dots dy_N,$$

$$\hat{h} := \hat{h}_{\text{LNO}}(r_1, r'_1, \dots, r_N, r'_N) := \frac{1}{n - N} \sum_{i \in [n] \setminus L(r_1, r_2, \dots, r_N)} k_\sigma(x_{i,1}, x_{r_1,r'_1}) \dots k_\sigma(x_{i,N}, x_{r_N,r'_N}),$$

192 where $L(r_1, r_2, \dots, r_N)$ is any subset of $[n]$ containing $\{r_1, r_2, \dots, r_N\}$ and having cardinality N .
 193 If r_1, r_2, \dots, r_N are not distinct, the additional indices can be chosen arbitrarily. For simplicity, we
 194 use a leave- N -out (LNO) estimator \hat{h}_{LNO} rather than a hybrid estimator like we used in the case of
 195 paired observations. As in the paired observation setting, we define

$$(\hat{w}, \hat{\alpha}) := \arg \min_{w \in \Delta^M, \alpha \in \Delta_M^{Nn}} \hat{J}(w, \alpha),$$

196 and similarly define $\hat{q} := q_{\hat{w}, \hat{\alpha}}$. Whenever \hat{w} , $\hat{\alpha}$, q , or $q_{\hat{w}, \hat{\alpha}}$ are referenced in the arbitrary group size
 197 setting, we will be referring to these estimators.

198 S3.2.2 Proof of Theorem 1: Arbitrary Group Size

199 We now state Theorem 1 for arbitrary group size.

200 **Theorem 1a.** *Given grouped observations of size N , let $\epsilon > 0$ and set $\delta =$
 201 $2(Nn)^N \exp \left\{ -\frac{\sigma^{2Nd} (n-N) \epsilon^2}{8C_k^{2N}} \right\}$. With probability at least $1 - \delta$ the following holds:*

$$\|q - q_{\hat{w}, \hat{\alpha}}\|_2^2 \leq \inf_{w \in \Delta^M, \alpha \in \Delta_M^{Nn}} \|q - q_{w, \alpha}\|_2^2 + \epsilon.$$

202 *Proof.* The proof proceeds as in the paired observation setting. In particular,

$$\begin{aligned} & P_q \left\{ \sup_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} |J(w, \alpha) - \hat{J}(w, \alpha)| > \frac{\epsilon}{2} \right\} \\ & \leq P_q \left\{ \sup_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \sum_{m,r,r'} \prod_{i \in [N]} w_m \alpha_{m,r_i,r'_i} \left| h(r_1, r'_1, \dots, r_N, r'_N) - \hat{h}(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{\epsilon}{4} \right\} \\ & \leq P_q \left\{ \max_{r_1, r'_1, \dots, r_N, r'_N} \left| h(r_1, r'_1, \dots, r_N, r'_N) - \hat{h}(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{\epsilon}{4} \right\} \\ & \leq \sum_{r,r'} P_q \left\{ \left| h(r_1, r'_1, \dots, r_N, r'_N) - \hat{h}(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{\epsilon}{4} \right\} \end{aligned}$$

203 The first step above is due to the triangle inequality, and the penultimate step is due to simplex con-
 204 straints on w, α . Let $k_i(r_1, r'_1, \dots, r_N, r'_N) := k_\sigma(x_{i,1}, x_{r_1, r'_1})k_\sigma(x_{i,2}, x_{r_2, r'_2}) \cdots k_\sigma(x_{i,N}, x_{r_N, r'_N})$.
 205 Noting that $h(r_1, r'_1, \dots, r_N, r'_N) = \mathbb{E}_q\{k_\sigma(x_{i,1}, x_{r_1, r'_1}) \cdots k_\sigma(x_{i,N}, x_{r_N, r'_N})\}$, we have

$$\begin{aligned} & P_q \left\{ \left| h(r_1, r'_1, \dots, r_N, r'_N) - \hat{h}(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{\epsilon}{4} \right\} \\ &= P_q \left\{ \left| \sum_{i \in [n] \setminus L(r_1, \dots, r_N)} \mathbb{E}_{(x_{i,1}, \dots, x_{i,N}) \sim q} \{k_i(r_1, r'_1, \dots, r_N, r'_N)\} \right. \right. \\ & \quad \left. \left. - k_i(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{(n-N)\epsilon}{4} \right\} \end{aligned}$$

206 The terms $k_i(r_1, r'_1, \dots, r_N, r'_N)$ are independent random variables due to use of the LNO estimator.
 207 By assumption, $0 \leq k_i(r_1, r'_1, \dots, r_N, r'_N) \leq C_k^N \sigma^{-Nd}$ so the k_i are bounded for fixed $\sigma > 0$. We
 208 apply Hoeffding's inequality

$$\begin{aligned} P_q \left\{ \left| h(r_1, r'_1, \dots, r_N, r'_N) - \hat{h}(r_1, r'_1, \dots, r_N, r'_N) \right| > \frac{\epsilon}{4} \right\} &\leq 2 \exp \left\{ -\frac{2(n-N)^2 \epsilon^2}{16(n-N) C_k^{2N} \sigma^{-2Nd}} \right\} \\ &= 2 \exp \left\{ -\frac{\sigma^{2Nd} (n-N) \epsilon^2}{8 C_k^{2N}} \right\}. \end{aligned}$$

209 Substituting backward we obtain the desired upper bound

$$\begin{aligned} P_q \left\{ \sup_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} |J(w, \alpha) - \hat{J}(w, \alpha)| > \frac{\epsilon}{2} \right\} &\leq \sum_{r_1, r'_1} \cdots \sum_{r_N, r'_N} 2 \exp \left\{ -\frac{\sigma^{2Nd} (n-N) \epsilon^2}{8 C_k^{2N}} \right\} \\ &= 2(Nn)^N \exp \left\{ -\frac{\sigma^{2Nd} (n-N) \epsilon^2}{8 C_k^{2N}} \right\} \end{aligned}$$

210 From here the proof is identical to the paired observation case, but with

$$\delta = 2(Nn)^N \exp \left\{ -\frac{\sigma^{2Nd} (n-N) \epsilon^2}{8 C_k^{2N}} \right\}.$$

211

□

212 S4 Theorem 2

213 In this section we give the proof of Theorem 2 for groups of size two, before extending it to groups
 214 of arbitrary size. For readability, we first present some intermediate results to be used in the main
 215 proofs.

216 S4.1 Intermediate Results

217 We first prove two supporting results.

218 **Lemma 1.** For any $1 \leq p < \infty$, any $f, g \in L^p$, and any integer $a \geq 2$,

$$\|f^{\times a} - g^{\times a}\|_p \leq \|f\|_p^{a-1} \|f - g\|_p + \|g\|_p \|f^{\times(a-1)} - g^{\times(a-1)}\|_p,$$

219 where $f^{\times a}(y_1, y_2, \dots, y_a) := f(y_1)f(y_2) \cdots f(y_a)$.

220 *Proof.* Let $f, g \in L^p$, $1 \leq p < \infty$. Then

$$\begin{aligned}
\|f^{\times a} - g^{\times a}\|_p &= \|f^{\times a} - f^{\times(a-1)} \times g + f^{\times(a-1)} \times g - g^{\times a}\|_p \\
&\leq \|f^{\times a} - f^{\times(a-1)} \times g\|_p + \|f^{\times(a-1)} \times g - g^{\times a}\|_p \\
&= \left(\int |f(x_1) \cdots f(x_{a-1})(f(x_a) - g(x_a))|^p dx_1 \dots dx_a \right)^{\frac{1}{p}} \\
&\quad + \left(\int |g(x_a)(f(x_1) \cdots f(x_{a-1}) - g(x_1) \cdots g(x_{a-1}))|^p dx_1 \dots dx_a \right)^{\frac{1}{p}} \\
&= \|f^{\times(a-1)}\|_p \|f - g\|_p + \|g\|_p \|f^{\times(a-1)} - g^{\times(a-1)}\|_p \\
&= \|f\|_p^{a-1} \|f - g\|_p + \|g\|_p \|f^{\times(a-1)} - g^{\times(a-1)}\|_p.
\end{aligned}$$

221

□

222 We have the following corollary.

223 **Corollary 1.** For any $1 \leq p < \infty$, any $f, g \in L^p$, and any integer $a \geq 2$,

$$\|f^{\times a} - g^{\times a}\|_p \leq \left(\sum_{b=1}^a \|f\|_p^{a-b} \|g\|_p^{b-1} \right) \|f - g\|_p.$$

224 *Proof.* The proof is by induction. Lemma 1 provides the base of the recursion for $a = 2$. Now
 225 suppose the statement is true for $a \geq 2$. To prove the statement for $a + 1$, we apply Lemma 1 again,
 226 together with the induction hypothesis, to get

$$\begin{aligned}
\|f^{\times(a+1)} - g^{\times(a+1)}\|_p &\leq \|f\|_p^a \|f - g\|_p + \|g\|_p \|f^{\times a} - g^{\times a}\|_p \\
&\leq \|f\|_p^a \|f - g\|_p + \|g\|_p \left(\sum_{b=1}^a \|f\|_p^{a-b} \|g\|_p^{b-1} \right) \|f - g\|_p \\
&= \left(\sum_{b=1}^{a+1} \|f\|_p^{a+1-b} \|g\|_p^{b-1} \right) \|f - g\|_p.
\end{aligned}$$

227 This completes the proof.

□

228 S4.2 Proof of Theorem 2: Groups of Size Two

229 We restate Theorem 2 for convenience.

230 **Theorem 2.** If $\sigma \rightarrow 0$ and $\frac{n\sigma^{4d}}{\log n} \rightarrow \infty$ as $n \rightarrow \infty$, then $\|q - q_{\hat{w}, \hat{\alpha}}\|_1 \xrightarrow{a.s.} 0$.

231 *Proof.* Lemma 3.1 of [3] states that if $\int \hat{q} = 1$ and $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$, then $\|\hat{q} - q\|_1 \xrightarrow{a.s.} 0$. Since
 232 $\int \hat{q} = 1$ in our case, our strategy is to show $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$. To do this it suffices to show that

$$\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w, \alpha}\|_2^2 \xrightarrow{a.s.} 0 \quad (\text{S5})$$

233 and

$$\inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w, \alpha}\|_2 \xrightarrow{a.s.} 0. \quad (\text{S6})$$

234 To show (S5), by the Borel-Cantelli lemma, it suffices to show that for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P_q \left(\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} \|q - q_{w, \alpha}\|_2^2 \geq \epsilon \right) < \infty.$$

235 Thus let $\epsilon > 0$. By Theorem 1a, the probability in question is at most

$$\begin{aligned}\delta &= 2(nN)^N \exp \left\{ -\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}} \right\} \\ &= 8 \exp \left\{ -2 \log n \left(\frac{(n-2)\sigma^{4d}\epsilon^2}{16C_k^4 \log n} - 1 \right) \right\}.\end{aligned}$$

236 By assumption on the growth of n and σ , there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\frac{(n-2)\sigma^{4d}\epsilon^2}{16C_k^4 \log n} \geq 2.$$

237 For such n we have

$$\delta \leq 8 \exp\{-2 \log n\} = \frac{8}{n^2}$$

238 which is summable.

239 To show (S6), let w^* be the true mixing weights from (2). For $i = 1, \dots, n$ let e_i be the $m \in [M]$

240 such that $X_i = (x_{i,1}, x_{i,2}) \stackrel{i.i.d.}{\sim} p_m^*$. Define

$$\begin{aligned}n_m &= |\{i : e_i = m\}|, & m &= 1, \dots, M \\ \alpha_{m,i,1}^* &= \alpha_{m,i,2}^* = \begin{cases} \frac{1}{2n_m}, & e_i = m \\ 0, & \text{otherwise} \end{cases}, & m &= 1, \dots, M\end{aligned}$$

241 With this ‘‘oracle’’ assignment of weights, $p(x; \alpha_m^*)$ is just the regular KDE for p_m^* . Therefore, we
242 may apply known results for consistency of standard KDEs. In particular, we will apply Theorem 3.1
243 of [3] which implies

$$\|p(\cdot; \alpha_m^*) - p_m^*\|_2 \xrightarrow{a.s.} 0 \text{ as } n_m \rightarrow \infty \quad (\text{S7})$$

244 provided $k \in L^2$ and $\sum_n \frac{1}{n^2 \sigma_n^d} < \infty$. Both of these conditions are satisfied by assumption in our
245 setting. Furthermore, as $n \rightarrow \infty$ we have $\frac{n_m}{n} \rightarrow w_m^*$ almost surely, and therefore $n_m \rightarrow \infty$ almost
246 surely.

247 Finally, we have

$$\begin{aligned}\inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{2n}}} \|q - q_{w,\alpha}\|_2 &\leq \|q - q_{w^*,\alpha^*}\|_2 \\ &= \left\| \sum_{m=1}^M w_m^* (p_m^* \times p_m^* - p(\cdot; \alpha_m^*) \times p(\cdot; \alpha_m^*)) \right\|_2 \\ &\leq \sum_{m=1}^M w_m^* \|p_m^* \times p_m^* - p(\cdot; \alpha_m^*) \times p(\cdot; \alpha_m^*)\|_2 \\ &\leq \sum_{m=1}^M w_m^* (\|p_m^*\|_2 + \|p(\cdot; \alpha_m^*)\|_2) \|p_m^* - p(\cdot; \alpha_m^*)\|_2 \\ &\leq \sum_{m=1}^M w_m^* 3 \|p_m^*\|_2 \|p_m^* - p(\cdot; \alpha_m^*)\|_2 \\ &\xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty,\end{aligned}$$

248 where the fourth step uses Lemma 1 and the fifth step holds for n sufficiently large (a.s.). This
249 completes the proof. \square

250 S4.3 Proof of Theorem 2: Arbitrary Group Size

251 We consider the problem for arbitrary group size as described in Section S3.2.1 of this document. The
252 proof of Theorem 2 for arbitrary group size is similar to the proof for groups of size two. The main
253 difference will be in use of Theorem 1a rather than Theorem 1 to invoke the Borel-Cantelli lemma.

254 **Theorem 2a.** Given grouped observations of size $N \in \mathbb{Z}^+$, if $\sigma \rightarrow 0$ and $\frac{n\sigma^{2Nd}}{\log n} \rightarrow \infty$ as $n \rightarrow \infty$
 255 then $\|q - \hat{q}\|_1 \xrightarrow{a.s.} 0$ as $n \rightarrow \infty$.

256 *Proof.* We will appeal to Lemma 3.1 of [3] as we did for groups of size two. Namely, if $\int \hat{q} = 1$ and
 257 $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$, then $\|\hat{q} - q\|_1 \xrightarrow{a.s.} 0$. Our strategy again is to show $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$. To do this it
 258 suffices to show that

$$\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2^2 \xrightarrow{a.s.} 0 \quad (\text{S8})$$

259 and

$$\inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2 \xrightarrow{a.s.} 0. \quad (\text{S9})$$

260 To show (S8), by the Borel-Cantelli lemma, it suffices to show that for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P_q \left(\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2^2 \geq \epsilon \right) < \infty.$$

261 Thus let $\epsilon > 0$. By Theorem 1a, the probability in question is at most

$$\begin{aligned} \delta &= 2(Nn)^N \exp\left\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\right\} \\ &= 2N^N \exp\left\{N \log n + \left(\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}}\right)\right\} \\ &= 2N^N \exp\left\{-N \log n \left(\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}N \log n} - 1\right)\right\}. \end{aligned}$$

262 By assumption on the growth of n and σ , there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}N \log n} \geq 2.$$

263 For such n we have

$$\delta \leq 2N^N \exp\{-N \log n\} = \frac{2N^N}{n^N}$$

264 which is summable for $N > 1$.

265 To show (S9), let w^* be the true mixing weights from (S4). For $i = 1, \dots, n$ let e_i be the $m \in [M]$
 266 such that $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,N}) \stackrel{i.i.d.}{\sim} p_m^*$. Define

$$\begin{aligned} n_m &= |\{i : e_i = m\}|, & m &= 1, \dots, M \\ \alpha_{m,i,j}^* &= \begin{cases} \frac{1}{n_m N}, & e_i = m \\ 0, & \text{otherwise} \end{cases}, & m &= 1, \dots, M, j = 1, \dots, N \end{aligned}$$

267 We are again using an ‘‘oracle’’ assignment of weights, so $p(x; \alpha_m^*)$ is just the regular KDE for p_m^* .
 268 Therefore, we may again apply Theorem 3.1 of [3] which implies

$$\|p(\cdot; \alpha_m^*) - p_m^*\|_2 \xrightarrow{a.s.} 0 \text{ as } n_m \rightarrow \infty \quad (\text{S10})$$

269 provided $k \in L^2$ and $\sum_n \frac{1}{n^2 \sigma_n^d} < \infty$. Both of these conditions are satisfied by assumption in our
 270 setting. Furthermore, as $n \rightarrow \infty$ we have $\frac{n_m}{n} \rightarrow w_m^*$ almost surely, and therefore $n_m \rightarrow \infty$ almost
 271 surely.

272 Finally, we have

$$\begin{aligned}
& \inf_{\substack{w \in \Delta_M^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2 \leq \|q - q_{w^*,\alpha^*}\|_2 \\
& = \left\| \sum_{m=1}^M w_m^* (p_m^{*\times N} - p(\cdot; \alpha_m^*)^{\times N}) \right\|_2 \\
& \leq \sum_{m=1}^M w_m^* \|p_m^{*\times N} - p(\cdot; \alpha_m^*)^{\times N}\|_2 \\
& \leq \sum_{m=1}^M w_m^* \left(\sum_{b=1}^N \|p_m^*\|_2^{N-b} \|p(\cdot; \alpha_m^*)\|_2^{b-1} \right) \|p_m^* - p(\cdot; \alpha_m^*)\|_2 \\
& \leq \sum_{m=1}^M w_m^* \left(\sum_{b=1}^N 2^{b-1} \|p_m^*\|_2^{N-1} \right) \|p_m^* - p(\cdot; \alpha_m^*)\|_2 \\
& \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty,
\end{aligned}$$

273 where the penultimate step uses (S10) and Corollary 1, and the final step holds for n sufficiently large
274 (a.s.). This completes the proof. \square

275 S5 Background on the Grouped Sample Setting and Proof of Theorem 3

276 Here we prove Theorem 3. We will be proving a general and more technical version of this theorem,
277 Theorem 5, from which Theorem 3 is a direct consequence. First we will introduce some background
278 to the problem setting which was introduced in [12]. **This section uses its own notation which does**
279 **not extend to other parts of the supplement or main text.**

280 S5.1 Identifiability in the Grouped Sample Setting

281 We will be concerned with probability measures on a measurable space (Ω, \mathcal{F}) . Let δ be the Dirac
282 measure. Let \mathcal{D} be the set of probability measures on (Ω, \mathcal{F}) . We call a probability measure on \mathcal{D} of
283 the form

$$\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$$

284 a *mixture of measures* [12]. For all mixtures of measures we will assume that $a_i > 0$ for all i and
285 $\mu_i \neq \mu_j$ when $i \neq j$ so that m is the number of distinct mixture components. The *grouped sample*
286 setting from [12] considers the situation where samples come in groups of size n by first sampling
287 a random measure component from a mixture of measures $\gamma \sim \mathcal{P}$, which is then sampled iid n
288 times. So one has access to samples of the form $\mathbf{X} = (X_1, \dots, X_n)$ with $X_1, \dots, X_n \stackrel{iid}{\sim} \gamma$. In this
289 situation the identifiability of \mathcal{P} depends on whether the distribution of \mathbf{X} is uniquely determined by
290 \mathcal{P} and the number of samples per group n . To this end [12] introduced the V_n operator which maps
291 a mixture of measures to the distribution of \mathbf{X} :

$$V_n \left(\sum_{i=1}^m a_i \delta_{\mu_i} \right) = \sum_{i=1}^m a_i \mu_i^{\times n},$$

292 where $\mu^{\times n}$ denotes the product measure n times. We note that $n = 1$ corresponds to a typical
293 mixture model where each mixture component is sampled once after being selected and there is no
294 grouped sample structure. For the grouped sample setting [12] introduces the following notion of
295 identifiability.

296 **Definition 1.** A mixture of measures, $\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$, is called n -identifiable if there does not
297 exist a different mixture of measures $\mathcal{Q} = \sum_{j=1}^{m'} b_j \delta_{\nu_j}$, with $m' \leq m$, such that $V_n(\mathcal{P}) = V_n(\mathcal{Q})$.

298 A *completely* rigorous mathematical treatment of the previous notions is a bit involved and can be
299 found in [12]. In [12] it is shown that if the mixture components are jointly irreducible then a mixture
300 of measures is 2-identifiable, if they are linearly independent then they are 3-identifiable, and that any
301 mixture of measures with m components is $(2m - 1)$ -identifiable.

302 S5.2 Notation

303 Before we state and prove the main theorem of this section we need to first introduce some notation.

304 Let S_m be the symmetric group over m symbols. Abusing notation slightly we will let the elements
305 of S_m be a group action on $[m]$ as well as \mathbb{R}^m . On \mathbb{R}^m it is defined as the following

$$\sigma \left([x_1, \dots, x_m]^T \right) = [x_{\sigma(1)}, \dots, x_{\sigma(m)}]^T. \quad (\text{S11})$$

306 We also let S_m be an operator where $S_m \cdot x$ is the orbit of x , i.e.

$$S_m \cdot x \triangleq \{ \sigma(x) : \sigma \in S_m \}. \quad (\text{S12})$$

307 Recall that for a pair of Hilbert spaces H, H' the direct sum $H \oplus H'$ is a Hilbert space with elements
308 of the form $x \oplus x'$ and inner product defined as $\langle x \oplus x', y \oplus y' \rangle = \langle x, y \rangle + \langle x', y' \rangle$. For a pair of
309 Banach spaces B, B' we define the direct sum via the norm $\|b \oplus b'\|_{B \oplus B'} \triangleq \|b\|_B + \|b'\|_{B'}$ which
310 is itself a Banach space ([1] p. 183).

311 For a pair of Hilbert spaces H, H' let $H \otimes H'$ be the tensor product of these two spaces and $h \otimes h'$
312 be the tensor product of vectors $h \in H$ and $h' \in H'$. For a vector in a Hilbert space h let $h^{\otimes n}$ denote
313 the tensor power, i.e. $\underbrace{h \otimes \dots \otimes h}_{n \text{ times}}$.

314 In the following the space of finite signed measures is equipped with the total variation topology and
315 unadorned norms refer to the total variation norm on finite signed measures, which forms a Banach
316 space. Norms for various Lebesgue spaces will have the associated subscript. Finally we note that
317 for two Hilbert spaces of square-integrable functions over σ -finite measure spaces $L^2(\Omega, \mathcal{F}, \mu)$ and
318 $L^2(\Omega', \mathcal{F}', \mu')$ we have that $L^2(\Omega, \mathcal{F}, \mu) \otimes L^2(\Omega', \mathcal{F}', \mu') \cong L^2(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mu \times \mu')$ via an
319 isomorphism $f \otimes f' \mapsto f \times f'$ ([4] Example 2.6.11) and we will use a 2 subscript for both norms.

320 S5.3 Full Theorem Statement and Proof

321 The following is the full general version of Theorem 3 and the main result of this section.

322 **Theorem 5.** *Let (Ω, \mathcal{F}) be a measurable space, $\mathcal{P} = \sum_{j=1}^m a_j \delta_{\mu_j}$ a mixture of measures on that
323 space which is n -identifiable, and $\mathcal{P}_i = \sum_{j=1}^{m'_i} b_{i,j} \delta_{\nu_{i,j}}$ a sequence of mixtures of measures with
324 $m'_i \leq m$ for all i , such that $V_n(\mathcal{P}_i) \rightarrow V_n(\mathcal{P})$. Then $m'_i \rightarrow m$ and there exists a sequence of
325 permutations σ_i such that $\sigma_i(b_i) \rightarrow a$ and $\nu_{i,\sigma_i(j)} \rightarrow \mu_j$ for all j .*

326 Essentially this says that as one finds grouped sample distributions $V_n(\mathcal{Q}_i)$ which approach the true
327 grouped sample distribution $V_n(\mathcal{P})$ the mixture of measures \mathcal{Q}_i will automatically recover the true
328 mixing weights and components from \mathcal{P} so long as \mathcal{P} is n -identifiable. In other words, one simply
329 needs to fit the grouped distribution $V_n(\mathcal{P})$ well to get a good estimate of the mixture components.
330 Theorem 3 from the main text is a direct consequence of Theorem 5.

331 **Corollary 2** (Theorem 3). *Let $\sum_{m=1}^M w_m p_m$ be an N -identifiable mixture
332 model, and $\sum_{m=1}^M \hat{w}_{m,j} \hat{p}_{m,j}$ be a sequence of mixture models such that
333 $\left\| \sum_{m=1}^M \hat{w}_{m,j} \hat{p}_{m,j}^{\times N} - \sum_{m=1}^M w_m p_m^{\times N} \right\|_1 \rightarrow 0$. Then there is a sequence of permutations σ_j
334 so that $\hat{w}_{\sigma_j(m),j} \rightarrow w_m$ and $\|\hat{p}_{\sigma_j(m),j} - p_m\|_1 \rightarrow 0$ for all m .*

335 We introduce some preliminary results before proving Theorem 5. The following lemma will be
336 needed for our proof.

337 **Lemma 2.** *Let \mathcal{P} and \mathcal{Q} be mixtures of measures, then $\|V_{n'}(\mathcal{P}) - V_{n'}(\mathcal{Q})\| \leq \|V_n(\mathcal{P}) - V_n(\mathcal{Q})\|$
338 for all $n' \leq n$.*

339 *Proof of Lemma 2.* From [2] (Section 3.1 Exercise 7a) we have the following

$$\begin{aligned}
& \|V_n(\mathcal{P}) - V_n(\mathcal{Q})\| \\
&= \sup \left\{ \sum_{i=1}^k |(V_n(\mathcal{P}) - V_n(\mathcal{Q}))(E_i)| : \right. \\
&\quad \left. k \in \mathbb{N}, E_1, \dots, E_k \in \mathcal{F}^{\times n} \text{ are disjoint, and } \bigcup_{i=1}^k E_i = \Omega^{\times n} \right\} \\
&\geq \sup \left\{ \sum_{i=1}^k |(V_n(\mathcal{P}) - V_n(\mathcal{Q}))(E_i \times \Omega^{\times n-n'})| : \right. \\
&\quad \left. k \in \mathbb{N}, E_1, \dots, E_k \in \mathcal{F}^{\times n'} \text{ are disjoint, and } \bigcup_{i=1}^k E_i = \Omega^{\times n'} \right\} \\
&= \sup \left\{ \sum_{i=1}^k |(V_{n'}(\mathcal{P}) - V_{n'}(\mathcal{Q}))(E_i)| : \right. \\
&\quad \left. k \in \mathbb{N}, E_1, \dots, E_k \in \mathcal{F}^{\times n'} \text{ are disjoint, and } \bigcup_{i=1}^k E_i = \Omega^{\times n'} \right\} \\
&= \|V_{n'}(\mathcal{P}) - V_{n'}(\mathcal{Q})\|.
\end{aligned}$$

340

□

341 The following lemma is the main workhorse in the proof of Theorem 5.

342 **Lemma 3.** Let (Ω, \mathcal{F}) be a measurable space, $\mathcal{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$ a mixture of measures on that
343 space, $n \in \mathbb{N}$, and $\mathcal{P}_i = \sum_{j=1}^{m'} b_j \delta_{\nu_{i,j}}$ a sequence of mixtures of measures (m' is fixed) with such
344 that $V_n(\mathcal{P}_i) \rightarrow V_n(\mathcal{P})$ (b does not depend on i). Then there exists a subsequence i_k and a collection
345 of probability measures $\nu_1, \dots, \nu_{m'}$ such that $\nu_{i_k,j} \rightarrow \nu_j$ for all j and $V_n(\mathcal{P}) = V_n(\sum_{j=1}^{m'} b_j \delta_{\nu_j})$.

346 *Proof of Lemma 3.* We will use bold symbols to represent elements that depend on i , e.g. $\nu_j = \nu_{i,j}$.
347 Let $\bar{\mu} = \sum_{k=1}^m a_k \mu_k$. By the Lebesgue-Radon-Nikodym Theorem ([2] Theorem 3.8) there exists
348 series of measures $\lambda_1, \dots, \lambda_{m'}$ and $\rho_1, \dots, \rho_{m'}$ such that $\nu_k = \lambda_k + \rho_k$ with $\lambda_k \perp \bar{\mu}$ and $\rho_k \ll \bar{\mu}$
349 for all $k \in [m']$.

350 For some fixed ℓ let \mathbf{A}_ℓ be the sequence of measurable sets such that $\lambda_\ell(\cdot \cap \mathbf{A}_\ell) = \lambda_\ell$ and $\bar{\mu}(\mathbf{A}_\ell) =$
351 0, this is possible since $\lambda_\ell \perp \bar{\mu}$. From Lemma 2 we have that

$$\left\| \sum_{k=1}^m a_k \mu_k - \sum_{j=1}^{m'} b_j \nu_j \right\| \rightarrow 0 \Rightarrow \left| \sum_{k=1}^m a_k \mu_k(\mathbf{A}_\ell) - \sum_{j=1}^{m'} b_j \nu_j(\mathbf{A}_\ell) \right| \rightarrow 0 \quad (\text{S13})$$

$$\Rightarrow \left| b_\ell \rho_\ell(\mathbf{A}_\ell) + b_\ell \lambda_\ell(\mathbf{A}_\ell) + \sum_{j \in [m'] \setminus \{\ell\}} b_j \nu_j(\mathbf{A}_\ell) \right| \rightarrow 0 \quad (\text{S14})$$

$$\Rightarrow \left| b_\ell \lambda_\ell(\mathbf{A}_\ell) + \sum_{j \in [m'] \setminus \{\ell\}} b_j \nu_j(\mathbf{A}_\ell) \right| \rightarrow 0. \quad (\text{S15})$$

352 Because all of the summands inside the absolute value on the last line are positive we have that
353 $\|\lambda_\ell\| \rightarrow 0$ and thus $\|\rho_\ell\| \rightarrow 1$. Eventually in our sequence we must have that $\|\rho_\ell\| > 0$, so eventually
354 in our subsequence we can define $\nu'_\ell = \rho_\ell / \|\rho_\ell\|$ which is now a sequence of probability measures
355 which are absolutely continuous with respect to $\bar{\mu}$ and $\|\nu'_\ell - \nu_\ell\| \rightarrow 0$.

356 From this we have that there exists sequences of probability measures $\nu'_1, \dots, \nu'_{m'}$ such that
357 $\|\nu_k - \nu'_k\| \rightarrow 0$ and $\nu'_k \ll \bar{\mu}$ for all $k \in [m']$. Lemma 3.3.7 in [11] states that, for probability mea-
358 sures over the same domain $\xi_1, \dots, \xi_d, \gamma_1, \dots, \gamma_d$ that $\left\| \prod_{j=1}^d \xi_j - \prod_{k=1}^d \gamma_k \right\| \leq \sum_{k=1}^d \|\xi_k - \gamma_k\|$.

359 It follows therefore that $\left\| \nu_k^{\times n} - \nu_k^{\times n} \right\| \rightarrow 0$ for all k and

$$\left\| \sum_{k=1}^m a_k \mu_k^{\times n} - \sum_{j=1}^{m'} b_j \nu_j^{\times n} \right\| \rightarrow 0. \quad (\text{S16})$$

360 For some fixed ℓ let \mathbf{q}'_ℓ be the Radon-Nikodym derivative of ν'_ℓ with respect to $\bar{\mu}$. Let $\mathbf{B}_\ell =$
 361 $\mathbf{q}'_\ell^{-1}([2/b_\ell, \infty))$. We have the following

$$\sum_{k=1}^{m'} b_k \nu'_k(\mathbf{B}_\ell) \geq b_\ell \nu'_\ell(\mathbf{B}_\ell) \quad (\text{S17})$$

$$\geq b_\ell \int_{\mathbf{B}_\ell} 2/b_\ell d\bar{\mu} \quad (\text{S18})$$

$$\geq 2\bar{\mu}(\mathbf{B}_\ell). \quad (\text{S19})$$

362 From Lemma 2 applied to (S16) we have that $\left| \sum_{k=1}^{m'} b_k \nu'_k(\mathbf{B}_\ell) - \bar{\mu}(\mathbf{B}_\ell) \right| \rightarrow 0$ and be-
 363 cause $\left| \sum_{k=1}^{m'} b_k \nu'_k(\mathbf{B}_\ell) - \bar{\mu}(\mathbf{B}_\ell) \right| \geq \bar{\mu}(\mathbf{B}_\ell)$ it follows that $\bar{\mu}(\mathbf{B}_\ell) \rightarrow 0$. Now we have that
 364 $\sum_{k=1}^{m'} b_k \nu'_k(\mathbf{B}_\ell) \rightarrow 0$ and thus $\nu'_\ell(\mathbf{B}_\ell) \rightarrow 0$.

365 Because $\nu'_\ell(\mathbf{q}'_\ell^{-1}([2/b_\ell, \infty))) \rightarrow 0$ and therefore $\nu'_\ell(\mathbf{B}_\ell^C) \rightarrow 1$, for sufficiently large i we can now
 366 define a sequence of probability measures ν''_ℓ via $\nu''_\ell(A) = \nu'_\ell(A \cap \mathbf{B}_\ell^C) / \nu'_\ell(\mathbf{B}_\ell^C)$. We have that

$$\|\nu'_\ell - \nu''_\ell\| = \|(\nu'_\ell(\mathbf{B}_\ell \cap \cdot) + \nu'_\ell(\mathbf{B}_\ell^C \cap \cdot)) - (\nu''_\ell(\mathbf{B}_\ell \cap \cdot) + \nu''_\ell(\mathbf{B}_\ell^C \cap \cdot))\| \quad (\text{S20})$$

$$\leq \|\nu'_\ell(\mathbf{B}_\ell \cap \cdot) - \nu''_\ell(\mathbf{B}_\ell \cap \cdot)\| + \|\nu'_\ell(\mathbf{B}_\ell^C \cap \cdot) - \nu''_\ell(\mathbf{B}_\ell^C \cap \cdot)\| \quad (\text{S21})$$

$$= \nu'_\ell(\mathbf{B}_\ell) + \|\nu'_\ell(\mathbf{B}_\ell^C \cap \cdot) - \nu''_\ell(\mathbf{B}_\ell^C \cap \cdot) / \nu'_\ell(\mathbf{B}_\ell^C)\| \quad (\text{S22})$$

$$= \nu'_\ell(\mathbf{B}_\ell) + |1 - 1/\nu'_\ell(\mathbf{B}_\ell^C)| \|\nu'_\ell(\mathbf{B}_\ell \cap \cdot)\| \quad (\text{S23})$$

367 which goes to zero, so $\|\nu''_\ell - \nu_\ell\| \rightarrow 0$. Note that ν''_ℓ is a sequence of probability measures with
 368 Radon-Nikodym derivatives $\mathbf{q}''_\ell \triangleq \mathbf{q}'_\ell \mathbf{1}_{\mathbf{B}_\ell^C} / \nu'_\ell(\mathbf{B}_\ell^C)$ ($\mathbf{1}$ is the indicator function) and thus

$$\sup_x \mathbf{q}''_\ell(x) = \sup_x \mathbf{q}'_\ell(x) \mathbf{1}_{\mathbf{B}_\ell^C}(x) / \nu'_\ell(\mathbf{B}_\ell^C) \leq 2/(b_\ell \nu'_\ell(\mathbf{B}_\ell^C))$$

and since $\nu_\ell(\mathbf{B}_\ell^C) \rightarrow 1$ eventually $\|\mathbf{q}''_\ell\|_\infty \leq 3/b_\ell$. From this we have that $\mathbf{q}''_\ell \in L^1(\Omega, \mathcal{F}, \bar{\mu}) \cap$
 $L^\infty(\Omega, \mathcal{F}, \bar{\mu})$ and $\|\mathbf{q}''_\ell\|_\infty$ is a bounded sequence. From Hölders's Inequality we have that

$$= \|\mathbf{q}''_\ell\|_2^2 = \|\mathbf{q}''_\ell \mathbf{q}''_\ell\|_1^2 \leq \|\mathbf{q}''_\ell\|_1 \|\mathbf{q}''_\ell\|_\infty = \|\mathbf{q}''_\ell\|_\infty$$

369 so \mathbf{q}''_ℓ is a bounded sequence in $L^2(\Omega, \mathcal{F}, \bar{\mu})$.

370 We now define $\nu''_1, \dots, \nu''_{m'}$, $\mathbf{q}''_1, \dots, \mathbf{q}''_{m'}$ similarly. There exists β such that $\|\mathbf{q}''_j\|_\infty \leq \beta$ and
 371 $\|\mathbf{q}''_j\|_2 \leq \beta$ along the whole series and for all j . Let $p_1, \dots, p_{m'}$ be the radon Nikodym derivatives
 372 for $\mu_1, \dots, \mu_{m'}$ with respect to $\bar{\mu}$, again these are in $L^1(\Omega, \mathcal{F}, \bar{\mu}) \cap L^2(\Omega, \mathcal{F}, \bar{\mu}) \cap L^\infty(\Omega, \mathcal{F}, \bar{\mu})$. To
 373 see this note that $p_i \leq 1/a_i$ otherwise we have that

$$\begin{aligned} \mu_i(p_i^{-1}((1/a_i, \infty))) &= \int_{p_i^{-1}((1/a_i, \infty))} p_i d\bar{\mu} \\ &> \int_{p_i^{-1}((1/a_i, \infty))} 1/a_i d\bar{\mu} \\ &> \sum_j \int_{p_i^{-1}((1/a_i, \infty))} 1/a_i a_j d\mu_j \\ &\geq \mu_i(p_i^{-1}((1/a_i, \infty))) \end{aligned}$$

374 a contradiction. Now we have

$$\left\| \sum_{k=1}^m a_k p_k^{\times n} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times n} \right\|_1 \rightarrow 0. \quad (\text{S24})$$

375 and Lemma 2 implies

$$\left\| \sum_{k=1}^m a_k p_k^{\times 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times 2} \right\|_1 \rightarrow 0. \quad (\text{S25})$$

376 From Hölder's Inequality ($\|f\|_2^2 \leq \|f\|_1 \|f\|_\infty$) we have that

$$\left\| \sum_{k=1}^m a_k p_k^{\times 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times 2} \right\|_2 \rightarrow 0 \quad (\text{S26})$$

377 and

$$\left\| \sum_{k=1}^m a_k p_k^{\otimes 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\otimes 2} \right\|_2 \rightarrow 0. \quad (\text{S27})$$

378 Let $S = \text{span}(\{p_1, \dots, p_m\})$ and $\ell \in [m']$ be arbitrary. We have that $\mathbf{q}_\ell'' = \text{proj}_S(\mathbf{q}_\ell'') +$
 379 $\text{proj}_{S^\perp}(\mathbf{q}_\ell'')$, noting that the summands in the decomposition are both L^2 bounded sequences. So
 380 now we have that

$$\left\langle \sum_{k=1}^{m'} b_k \mathbf{q}_k''^{\otimes 2} - \sum_{j=1}^m a_j p_j^{\otimes 2}, \text{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \right\rangle \rightarrow 0 \quad (\text{S28})$$

$$\Rightarrow \left\langle \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\otimes 2}, \text{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \right\rangle \rightarrow 0 \quad (\text{S29})$$

$$\Rightarrow b_\ell \langle \text{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2}, \text{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \rangle + \sum_{j \in [m'] \setminus \{\ell\}} b_j \langle \mathbf{q}_j'', \text{proj}_{S^\perp}(\mathbf{q}_j'') \rangle^2 \rightarrow 0 \quad (\text{S30})$$

$$\Rightarrow b_\ell \|\text{proj}_{S^\perp}(\mathbf{q}_\ell'')\|_2^4 \rightarrow 0. \quad (\text{S31})$$

381 From this we have that $\|\text{proj}_S(\mathbf{q}_k'') - \mathbf{q}_k''\|_2 \rightarrow 0$ for all k . Since $\bigoplus_{j=1}^{m'} \text{proj}_S(\mathbf{q}_j'')$ is a L^2 bounded
 382 sequence on a finite dimensional space by the Bolzano-Weierstrass theorem it has a convergent
 383 subsequence which converges to $\bigoplus_{j=1}^{m'} q_j''$ so $\mathbf{q}_j'' \rightarrow q_j''$ in L^2 . From Hölder's Inequality we have
 384 that, along this subsequence

$$\|\mathbf{q}_k'' - q_k''\|_1 \leq \|\mathbf{q}_k'' - q_k''\|_2 \|1\|_2 \leq \|\mathbf{q}_k'' - q_k''\|_2 \sqrt{\int 1^2 d\bar{\mu}} = \|\mathbf{q}_k'' - q_k''\|_2 \rightarrow 0 \quad (\text{S32})$$

385 so q_k'' is a probability density for all k , since they must be nonnegative to converge and integrate to
 386 one. Now we have that

$$\sum_{j=1}^m a_j p_j^{\times n} = \sum_{k=1}^{m'} b_k q_k''^{\times n}. \quad (\text{S33})$$

387 And defining ν_k as the probability measure associated with q_k'' we have that there exists a subsequence
 388 such that $\|\nu_k - \nu_k\| \rightarrow 0$ for all k and

$$\sum_{j=1}^m a_j \mu_j^{\times n} = \sum_{k=1}^{m'} b_k \nu_k^{\times n}. \quad (\text{S34})$$

389

□

390 We can now prove Theorem 5.

391 *Proof of Theorem 5.* To help lighten notation we will simply bold some elements which depend
 392 on the sequence \mathcal{P}_i . Let $\mathcal{P}_i = \sum_{j=1}^{m'_i} \mathbf{b}_j \delta_{\nu_j}$ be a sequence of mixtures of measures (\mathbf{b}_j, ν_j are
 393 functions of i) such that $V_n(\mathcal{P}_i) \rightarrow V_n(\mathcal{P})$.

394 We define $\tilde{\mathbf{b}}$ a sequence in Δ^m so that $\tilde{\mathbf{b}}_j = \mathbf{b}_j$ for $j \leq m'_i$ and $\tilde{\mathbf{b}}_k = 0$ for $k > m'_i$. Consider the
 395 case where there exists no sequence of permutations such that $\sigma(\tilde{\mathbf{b}}) \rightarrow a$. From this it would follow
 396 that there exists a subsequence on i and $\varepsilon > 0$ such that $\|\tilde{\mathbf{b}} - \sigma(a)\| > \varepsilon$ for all $\sigma \in S_m$. The space

$$\Delta^m \cap \left(\bigcap_{\sigma \in S_m} \text{ball}(\sigma(a), \varepsilon)^C \right) \quad (\text{S35})$$

397 is compact (the ball is open) so there exists a sub-subsequence of i where $\tilde{\mathbf{b}}$ converges to a point
 398 $b \notin S_m \cdot a$. Let $I \subset [m]$ be the indices of b which are nonzero and $m' = \max(I)$. For sufficiently
 399 large i along our sub-subsequence we have that $m'_i \geq m'$ and furthermore

$$\left\| \sum_{j=1}^{m'_i} \mathbf{b}_j \nu_j^{\times n} - \sum_{k \in I} b_k \nu_k^{\times n} \right\| \leq \left\| \sum_{j \in I} \mathbf{b}_j \nu_j^{\times n} - \sum_{k \in I} b_k \nu_k^{\times n} \right\| + \left\| \sum_{j \in I^C} \mathbf{b}_j \nu_j^{\times n} \right\| \quad (\text{S36})$$

$$\leq \left\| \sum_{j \in I} (\mathbf{b}_j - b_j) \nu_j^{\times n} \right\| + \left| \sum_{j \in I^C} \mathbf{b}_j \right| \quad (\text{S37})$$

$$\leq \sum_{j \in I} |(\mathbf{b}_j - b_j)| + \left| \sum_{j \in I^C} \mathbf{b}_j \right| \rightarrow 0 \quad (\text{S38})$$

400 and therefore

$$\left\| \sum_{k=1}^m a_k \mu_k^{\times n} - \sum_{j \in I} b_j \nu_j^{\times n} \right\| \rightarrow 0. \quad (\text{S39})$$

401 From Lemma 3 we have that there exists a subsequence of this sub-subsequence such that for $k \in I$
 402 there exists probability measures ν_k with $\|\nu_k - \mu_k\| \rightarrow 0$ and

$$\sum_{k=1}^m a_k \mu_k^{\times n} = \sum_{j \in I} b_j \nu_j^{\times n}. \quad (\text{S40})$$

403 If $|I| < m$ or $\nu_j = \nu_k$ for any $k \neq j$ and $j, k \in I$ then we have clearly violated identifiability since
 404 we can construct a mixture of measures \mathcal{P}' with fewer components than \mathcal{P} and $V_n(\mathcal{P}') = V_n(\mathcal{P})$.
 405 If $|I| = m$ (i.e. $I = [m]$) and ν_j are all distinct we have also arrived at a contradiction since letting
 406 $\mathcal{P}' = \sum_{j=1}^m b_j \delta_{\nu_j} \neq \mathcal{P}$ because there exists no σ such that $\sigma(b) = a$ and $V_n(\mathcal{P}') = V_n(\mathcal{P})$,
 407 contradicting identifiability.

408 So we have that for sufficiently large i that $m'_i = m$ and there exists at least one sequence σ such
 409 that $\sigma(\mathbf{b}) \rightarrow a$. So let $\left\| \sum_{k=1}^m a_k \mu_k^{\times n} - \sum_{j=1}^m \mathbf{b}_j \nu_j^{\times n} \right\| \rightarrow 0$. From what we have just shown, we
 410 can permute the indices and, without loss of generality, we can assume that $\mathbf{b} \rightarrow a$. So now we have
 411 that $\left\| \sum_{i=1}^m a_i \mu_i^{\times n} - \sum_{j=1}^m a_j \nu_j^{\times n} \right\| \rightarrow 0$.

Let $\tilde{S}_m \subset S_m$ be the subgroup of permutations such that $\sigma(a) = a$ for $\sigma \in \tilde{S}_m$ (also known as the
 stabilizer of a). Note that if a_1, \dots, a_m are distinct then \tilde{S}_m only contains the identity. We proceed
 by contradiction: suppose there exists no sequence of permutations $\sigma \in \tilde{S}_m$ such that $\nu_{\sigma(k)} \rightarrow \mu_k$
 for all k . From this it follows that there exists a subsequence and a $\varepsilon > 0$, such that $\bigoplus_{k=1}^m \nu_k$
 does not lie in $\bigcap_{\sigma \in \tilde{S}_m} (\text{ball}(\bigoplus_{k=1}^m \mu_{\sigma(k)}, \varepsilon)^C$. From Lemma 3 there exists probability measures,

ν_1, \dots, ν_m such that for some subsequence $\|\nu_k - \nu_k\| \rightarrow 0$ for all k and

$$\sum_{j=1}^m a_j \mu_j^{\times n} = \sum_{k=1}^m a_k \nu_k^{\times n}.$$

412 Because $\bigcap_{\sigma \in \tilde{S}_m} (\text{ball}(\bigoplus_{k=1}^m \mu_{\sigma(k)}, \varepsilon))^C$ is closed we have $\bigoplus_{j=1}^m \nu_j \in$
 413 $\bigcap_{\sigma \in \tilde{S}_m} (\text{ball}(\bigoplus_{k=1}^m \mu_{\sigma(k)}, \varepsilon))^C$ and there exists no $\sigma \in \tilde{S}_m$ such that $\nu_{\sigma(k)} = \mu_k$ for all
 414 k so. Setting $\mathcal{P}' = \sum_{k=1}^m a_k \delta_{\nu_k}$ we have that $\mathcal{P}' \neq \mathcal{P}$ but $V_n(\mathcal{P}') = V_n(\mathcal{P})$, a contradiction.
 415 □

416 S6 General Version of Corollary 1

417 Here we present the general version of Corollary 1 which guarantees recovery of the true mixture
 418 components using our estimator for any mixture model, provided there are a sufficient number of
 419 samples per group. For a mixture model $p = \sum_{m=1}^M w_m^* p_m^*$, using the estimator \hat{q} from Section
 420 S3.2.1 to estimate (S4):

$$q(y_1, y_2, \dots, y_N) = \sum_{m=1}^M w_m^* p_m^*(y_1) p_m^*(y_2) \dots p_m^*(y_N), \quad y_1, y_2, \dots, y_N \in \mathbb{R}^d.$$

421 combining Theorem 2a and Theorem 3 gives the following result.

422 **Corollary 3.** If $\sigma \rightarrow 0$ and $\frac{n\sigma^{2Nd}}{\log n} \rightarrow \infty$ as $n \rightarrow \infty$, and p is N -identifiable (e.g. $N = 2M - 1$),
 423 then $\hat{w}_m \xrightarrow{a.s.} w_m^*$ and $\|p(\cdot; \hat{\alpha}_m) - p_m^*\|_1 \xrightarrow{a.s.} 0$, up to a permutation.

424 References

- 425 [1] Stefan Banach. *Théorie des opérations linéaires*. 1932.
- 426 [2] Gerald B. Folland. *Real analysis: modern techniques and their applications*. Pure and applied
 427 mathematics. Wiley, 1999.
- 428 [3] L. Györfi and E. Masry. The L_1 and L_2 strong consistency of recursive kernel density estimation
 429 from dependent samples. *IEEE Transactions on Information Theory*, 36(3):531–539, 1990.
- 430 [4] R.V. Kadison and J.R. Ringrose. *Fundamentals of the theory of operator algebras. VI: Element-*
 431 *ary theory*. Pure and Applied Mathematics. Elsevier Science, 1983.
- 432 [5] JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *Journal of Machine*
 433 *Learning Research*, 13(Sep):2529–2565, 2012.
- 434 [6] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling
 435 for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM*
 436 *SIGIR conference on Research and Development in Information Retrieval*, pages 165–174,
 437 2016.
- 438 [7] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of
 439 topic coherence. In *Human language technologies: The 2010 annual conference of the North*
 440 *American chapter of the association for computational linguistics*, pages 100–108. Association
 441 for Computational Linguistics, 2010.
- 442 [8] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models
 443 with latent feature word representations. *Transactions of the Association for Computational*
 444 *Linguistics*, 3:299–313, 2015.
- 445 [9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for
 446 word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages
 447 1532–1543, 2014.
- 448 [10] Jipeng Qiang, Qian Zhenyu, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling
 449 techniques, applications, and performance: A survey. *arXiv preprint arXiv:1904.07695*, 2019.

- 450 [11] R.D. Reiss. *Approximate distributions of order statistics: with applications to nonparametric*
451 *statistics*. Springer series in statistics. Springer, 1989.
- 452 [12] Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric
453 mixture models. *The Annals of Statistics*, 47(5):2704–2733, October 2019.