Journal of Machine Learning Research 20 (2019) 1-57 Submitted 9/17; Revised 11/18; Published 1/19

Decontamination of Mutual Contamination Models

Julian Katz-Samuels JKATZSAM@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122 USA

Gilles Blanchard GILLES.BLANCHARD@MATH.UNI-POTSDAM.DE
Universität Potsdam, Institut für Mathematik
D-14476 Potsdam, Germany

Clayton Scott CLAYSCOT@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122 USA

Editor: Inderjit Dhillon

Abstract

Many machine learning problems can be characterized by *mutual contamination models*. In these problems, one observes several random samples from different convex combinations of a set of unknown base distributions and the goal is to infer these base distributions. This paper considers the general setting where the base distributions are defined on arbitrary probability spaces. We examine three popular machine learning problems that arise in this general setting: multiclass classification with label noise, demixing of mixed membership models, and classification with partial labels. In each case, we give sufficient conditions for identifiability and present algorithms for the infinite and finite sample settings, with associated performance guarantees.

Keywords: multiclass classification with label noise, classification with partial labels, mixed membership models, topic modeling, mutual contamination models

1. Introduction

In many machine learning problems, the learner observes several random samples from different mixtures of unknown base distributions, with unknown mixing weights, and the goal is to infer these base distributions. Examples include binary classification with label noise, multiclass classification with label noise, classification with partial labels, and topic modeling. The goal of this paper is to develop a unified framework and set of tools to study statistical properties of these problems in a very general setting.

To this end, we use the general framework of mutual contamination models (Blanchard and Scott, 2014). In a mutual contamination model, there are L distributions $P_1, \ldots, P_L$ called *base distributions*. The learner observes M random samples

$$X_1^i, \ldots, X_{n_i}^i \overset{i.i.d.}{\sim} \tilde{P}_i = \sum_{j=1}^{L} \pi_{i,j} P_j \tag{1}$$

where $i = 1, \ldots, M$, $\pi_{i,j} \geqslant 0$, and $\sum_j \pi_{i,j} = 1$. Here $\pi_{i,j}$ is the probability that an instance of the *contaminated distribution* $\tilde{P}_i$ is a realization of P_j. The $\pi_{i,j}$s and P_js are unknown and the $\tilde{P}_i$s are observed through data. In this work, we avoid parametric models and assume that the sample space is arbitrary. The model can be stated concisely as

$$\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}\boldsymbol{P} \tag{2}$$

where $\boldsymbol{P} = (P_1, \ldots, P_L)^T$, $\tilde{\boldsymbol{P}} = (\tilde{P}_1, \ldots, \tilde{P}_M)^T$, and $\boldsymbol{\Pi} = \left(\pi_{i,j}\right)$ is an $M \times L$ matrix that we call the *mixing matrix*.

In this paper we study *decontamination* of mutual contamination models, which is the problem of recovering, or estimating, the base distributions $\boldsymbol{P}$ from the contaminated distributions $\tilde{\boldsymbol{P}}$ from which data are observed, without knowledge of the mixing matrix $\boldsymbol{\Pi}$. We focus our attention on three specific types of mutual contamination models, all of which describe modern problems in machine learning: multiclass classification with label noise, demixing of mixed membership models and classification with partial labels. We will demonstrate that these three decontamination problems can be addressed using a common set of concepts and techniques. Before elaborating our contributions in detail, we first offer an overview of the three specific mutual contamination models, and associated decontamination problems, that we study.

Multiclass Classification with Label Noise: In multiclass classification with label noise, $M = L$ and the goal is to recover $\boldsymbol{P}$. Each P_i represents the distribution of a class of examples. The learner observes training examples with noisy labels, that is, realizations from the $\tilde{P}_j$s. This problem arises in nuclear particle classification (Scott et al., 2013). When one draws samples of a specific particle, it is impossible to remove other types of particles from the background. Thus, each example is drawn from a mixture of the different types of particles.

Demixing of Mixed Membership Models: We consider the following decontamination problem in mixed membership models: given a sample from each $\tilde{P}_i$, recover $\boldsymbol{P}$ up to a permutation. We refer to this decontamination problem as *demixing of mixed membership models*. This problem arises in the task of automatically uncovering the thematic topics of a corpus of documents. Under the mixed membership model approach, the words of each document are thought of as being drawn from a document-specific mixture of topics. Specifically, documents correspond to the $\tilde{P}_i$s and the topics to the P_is. This approach is also referred to as topic modeling. As we discuss in the next section, our theory significantly generalizes existing topic modeling theoretical guarantees.

Classification with Partial Labels:[1] In classification with partial labels, each data point is labeled with a *partial label* $Y \subset \{1, \ldots, L\}$; the true label is in Y, but it is not known which label is the true one. In our setup, we view the ith random sample as having partial label $Y_i \coloneqq \{j : \boldsymbol{\pi}_{i,j} > 0\}$ and being distributed according to $\tilde{P}_i = \sum_{j \in Y_i} \pi_{i,j} P_j$. Thus, the learner observes training examples from the contaminated distributions $\tilde{\boldsymbol{P}}$ and the *partial label matrix* $\boldsymbol{\Pi}^+ = (\mathbf{1}_{\{\boldsymbol{\Pi}_{i,j} > 0\}})$, and the goal is to recover $\boldsymbol{P}$.

There are many applications of classification with partial labels because often abundant sources of data are naturally associated with information that can be interpreted as partial

1. Classification with partial labels has also been referred to as the "superset learning problem" or the "multiple label problem" (Liu and Dietterich, 2014).

labels. For example, consider the task of face recognition. On the internet, there are many images with captions that indicate who is in the picture but do not indicate which face belongs to which person. A partial label could be formed by associating each face with the names of the individuals appearing in the same image (Cour et al., 2011).

Although our work emphasizes recovery of $\boldsymbol{P}$, it is also possible to think of decontamination of mutual contamination models as concerned with estimation of the mixing matrix $\boldsymbol{\Pi}$. This estimate of $\boldsymbol{\Pi}$ could be used as a plug-in for recently developed debiased losses for multiclass classification with label noise and classification with partial labels, which require knowledge of $\boldsymbol{\Pi}$ (Cid-Sueiro, 2012; Menon et al., 2015b; van Rooyen and Williamson, 2015; Patrini et al., 2017).

In this paper, we make the following contributions: (i) We give sufficient conditions on $\boldsymbol{P}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Pi}^+$ for identifiability of the three problems. (ii) We establish necessary conditions that in some cases match or are similar to the sufficient conditions. (iii) We introduce novel algorithms for the infinite and finite sample settings. These algorithms are nonparametric in the sense that they do not model P_i as a probability vector or other parametric model. Our algorithmic contributions show that while all three problems can be described in a unified way, the special structure of multiclass classification with label noise allows for a substantially simpler algorithm. (iv) We develop novel estimators for distributions obtained by iteratively applying the κ^* operator (defined below). (v) Finally, our framework gives rise to several novel geometric insights about each of these three problems and leverages concepts from affine geometry, multilinear algebra, and probability.

1.1. Notation

Let $\mathbb{Z}^+$ denote the positive integers. For $n \in \mathbb{Z}^+$, let $[n] = \{1, \ldots, n\}$. If $\boldsymbol{x} \in \mathbb{R}^K$, let x_i denote the ith entry of $\boldsymbol{x}$. If $\boldsymbol{x}_j \in \mathbb{R}^K$, then $x_{j,i}$ denotes the ith entry of $\boldsymbol{x}_j$. Let $\boldsymbol{e}_i$ denote the length L vector with 1 in the ith position and zeros elsewhere. Let $\boldsymbol{\pi}_i \in \Delta_L \subset \mathbb{R}^L$ be the transpose of the ith row of $\boldsymbol{\Pi}$ where Δ_L denotes the $(L-1)$-dimensional simplex, i.e., $\Delta_L = \{\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T \in \mathbb{R}^L \mid \sum_{i=1}^L \mu_i = 1 \text{ and } \forall i : \mu_i \geqslant 0\}$. Let Δ_L^M denote the product of M $(L-1)$-dimensional simplices, viewed as the space of $M \times L$ row-stochastic matrices. Let $\mathcal{P}$ denote the space of probability distributions on a measurable space $(\mathcal{X}, \mathcal{C})$. Let $\text{supp}(F)$ denote the support of a distribution F on a Borel space.

2. Related Work

Our work makes various contributions to the statistical understanding of multiclass classification with label noise, demixing of mixed membership models, and classification with partial labels. In the following subsections, we discuss how our results improve upon and relate to previous results in the literature.

2.1. Multiclass Classification with Label Noise

There has not been much work on classification with multiclass label noise. By contrast, label noise in the binary setting has received a fair amount of attention. For a review of work prior to 2013, see Scott et al. (2013). More recently, Natarajan et al. (2013) considered the binary label noise case where the label noise rates are known (in our case, the label noise

rates are unknown). van Rooyen and Williamson (2015) generalized the work of Natarajan et al. (2013) to the multiclass case, but again assumed that the mixing proportions are known. Recent work has proposed various algorithms for the binary setting where the label noise rates are unknown (Scott, 2015; van Rooyen et al., 2015; Menon et al., 2015a), but these algorithms have not been generalized to the multiclass case. Menon et al. (2016) consider the binary setting with instance-dependent corruption, but they assume that the class probability functions take the form of a single-index model, whereas we make no parametric assumptions on the P_is. Ghosh et al. (2017) consider multiclass label noise, but they make two restrictive assumptions: *(i)* in the infinite sample setting, they assume that there exists some function belonging to the chosen hypothesis class that attains 0 risk and *(ii)* in the finite sample setting, they assume that the label noise is symmetric, i.e., there exists a constant $c \in (0,1)$ such that $\pi_{i,j} = \frac{c}{L-1}$ for all $i \neq j$. Patrini et al. (2017) also study the multiclass setting, but they assume that if their neural network has access to sufficiently many samples, it can perfectly model $\Pr(\tilde{Y} = k \,|\, \boldsymbol{x})$ where $\boldsymbol{x}$ is a given feature vector and $\tilde{Y}$ is a corrupted label. Unlike most previous work that aims to learn a classifier, our focus is on estimating the base distributions. Given these estimates, one could then design a classifier to optimize some performance measure. See, for example, Section 4.3 of our initial work on this subject (Blanchard and Scott, 2014).

Another approach for modeling random label noise, in addition to the mutual contamination model, is the label flipping model. Indeed, several of the above-cited papers adopt this setting. In this model, the label Y of a data point is flipped independently of its features X and

$$\mu_{l,k} \coloneqq \Pr(\tilde{Y} = k \,|\, Y = l)$$

gives the probability that a data point with true label $Y = l$ is corrupted to have an observed label $\tilde{Y} = k$. Under the assumption that Y and X are jointly distributed, the $\mu_{l,k}$s can be related to the $\pi_{i,j}s$ via Bayes' rule. We choose to study the mutual contamination model because we find it more convenient to study the question of identifiability.

In this paper, we extend Scott et al. (2013), which examined binary classification with label noise (the case where $M = L = 2$). The multiclass setting is significantly more challenging and, as such, requires novel sufficient conditions and mathematical notions. In particular, Scott et al. (2013) use the notion of *irreducibility* of distributions as one of their sufficient conditions.

Definition 1 *For distributions G and H, we say that G is* irreducible *with respect to H if it is not possible to write $G = \gamma H + (1-\gamma)F$ where F is a distribution and $0 < \gamma \leqslant 1$.*

Definition 2 *For distributions G and H, we say that G and H are* mutually irreducible *if G is irreducible with respect to H and H is irreducible with respect to G. We denote*

$$IR = \{(G,H) : G \text{ and } H \text{ are mutually irreducible distributions}\}.$$

Scott et al. (2013) require that P_1 and P_2 are mutually irreducible. To treat the multiclass setting, we introduce a generalization of mutual irreducibility, namely *joint irreducibility*.

The work presented below on multiclass label noise originally appeared in a conference paper (Blanchard and Scott, 2014). The purpose of the present paper is to demonstrate that

the framework developed in that paper can be extended to the other two decontamination problems, and to provide a unified presentation of the three settings. In particular, the joint irreducibility assumption plays a pivotal role in all three settings, as does the task of mixture proportion estimation. However, the decontamination procedures for the latter two problems are substantially more complicated than for multiclass classification with label noise.

2.2. Demixing Mixed Membership Models

Mixed membership models have become a powerful modeling tool for data where data points are associated with multiple distributions. Applications have appeared in a wide range of fields including image processing (Li and Perona, 2005), population genetics (Pritchard et al., 2000), document analysis (Blei et al., 2003), and surveys (Berkman et al., 1989). One particularly popular application is topic modeling on a corpus of documents, such as the articles published in the journal Science. Topic modeling is closely related to demixing of mixed membership models and our work may be viewed as studying topic modeling on general domains.

In topic modeling, the base distributions P_i correspond to topics and the contaminated distributions $\tilde{P}_i$ to documents, which are regarded as mixtures of topics. In most cases, the P_is are assumed to have a finite sample space. A variety of approaches have been proposed for topic modeling. The most common approach assumes a generative model for a corpus of documents and determines the maximum likelihood fit of the model given data. However, because maximum likelihood is NP-hard, these approaches must rely on heuristics that can get stuck in local minima (Arora et al., 2012).

Recently, a trend towards algorithms for topic modeling with provable guarantees has emerged. Most of these methods rely on the separability assumption **(SEP)** and its variants (Donoho and Stodden, 2003; Arora et al., 2012, 2013; Ding et al., 2013, 2014; Recht et al., 2012; Huang et al., 2016). According to **(SEP)**, $P_1, \ldots, P_L$ are distributions on a finite sample space and for every $i \in \{1, \ldots, L\}$, there exists a word $x \in \text{supp}(P_i)$ such that $x \notin \cup_{j \neq i} \text{supp}(P_j)$. Our requirement that $P_1, \ldots, P_L$ are jointly irreducible is a natural generalization of separability of $P_1, \ldots, P_L$, as we will argue below. Specifically, if $P_1, \ldots, P_L$ have discrete sample spaces, separability and joint irreducibility coincide; however, if $P_1, \ldots, P_L$ are continuous, under joint irreducibility, $P_1, \ldots, P_L$ can have the same support.

A key ingredient in these algorithms is to use the assumption of a finite sample space to view the distributions as probability vectors in Euclidean space; this leads to approaches based on non-negative matrix factorization (NMF), linear programs, and random projections (Donoho and Stodden, 2003; Arora et al., 2012, 2013; Ding et al., 2013, 2014; Recht et al., 2012; Huang et al., 2016). However, more general distributions cannot be viewed as finite-dimensional vectors. Therefore, topic modeling on general domains requires new techniques. Our work seeks to provide such techniques.

Topic modeling on general domains has several applications, including in high-energy physics (Metodiev and Thaler, 2018a,b). In collider data, quantum chromodynamics causes data samples to be a mixture of different types of particles, where the underlying fraction of the particle type is unknown. In this setting, it is of interest to recover information about

each of the particles. Recently, Metodiev and Thaler (2018b) applied the Demix algorithm, Algorithm 4 in the current paper, to this problem in the case $M = L = 2$.

Topic modeling on general domains is also relevant to recent empirical research on topic modeling with word embeddings, e.g., (Das et al., 2015; Li et al., 2016b,a; Xun et al., 2017; Zhao et al., 2018). Word embeddings map words to vectors in $\mathbb{R}^d$ in a semantically and syntactically meaningful way. Their use has been pivotal to the state-of-art performance of many algorithms in NLP (Luong et al., 2013). Several algorithms for topic modeling with word embeddings model the topics as multivariate Gaussian distributions in order to handle words that do not belong to the vocabulary of the training dataset (Das et al., 2015; Xun et al., 2017). Whereas current topic modeling algorithms with theoretical guarantees do not cover such a modeling approach, the generality of our algorithms does.

2.3. Classification with Partial Labels

Classification with partial labels has had two main formulations in previous work (Liu and Dietterich, 2014). In one formulation **(PL-1)**, instances from each class are drawn independently and the partial label for each instance is drawn independently from a set-valued distribution. In another formulation **(PL-2)**, training data are in the form of bags where each bag is a set of instances and the bag has a set of labels. Each instance belongs to a single class, and the set of labels associated with the bag is given by the union of the labels of the instances in the bag. Our framework is similar to **(PL-2)**, although it does not assume a joint distribution on the features of instances and the partial labels.

Most work takes an empirical risk minimization approach to classification with partial labels (Jin and Ghahramani, 2002; Nyugen and Caruana, 2008; Cour et al., 2011; Liu and Dietterich, 2012). Typically, these algorithms aim to pick a classifier that minimizes the *partial label error*: the probability that a given classifier assigns a label to a training instance that is not contained in the partial label associated with the training instance. By contrast, our approach is to estimate the base distributions. One could then use these estimates to train a classifier under some performance measure.

There has not been much theoretical work on developing a statistical understanding of classification with partial labels. Cid-Sueiro (2012) and van Rooyen and Williamson (2015) develop methods for classification with partial labels that require knowledge of the mixing proportions, e.g., the probability that a label is in a partial label, given the true label. In this work, we make the more realistic assumption that the mixing proportions are unknown.

Liu and Dietterich (2014) consider the question of learnability where the mixing proportions are unknown. They consider two main sufficient conditions for learnability of a partial label problem. First, they require that for every label $l \in [L]$, the probability that l occurs with any particular distinct label l' is less than 1. Our condition on the partial label (described in the next Section) is considerably weaker. For example, it permits the case where there are two labels $l \neq l'$ such that whenever l occurs in a partial label, l' also occurs.

The second sufficient condition of Liu and Dietterich (2014) is based on the class distributions, partial label distributions *and* the hypothesis class of choice. It requires that every hypothesis that attains zero partial label error also attains zero true error. While this condition may be useful for the selection of a suitable hypothesis class for an ERM

approach, it is important to develop interpretable sufficient conditions that only depend on the characteristics of a partial label problem. Our work provides such conditions.

We also note that Liu and Dietterich (2014) consider the realizable case, that is, the case where the supports of $P_1, \ldots, P_L$ do not overlap. By contrast, we make the significantly weaker assumption that $P_1, \ldots, P_L$ are jointly irreducible, which allows $P_1, \ldots, P_L$ to have the same support. Thus, our work addresses the agnostic case in classification with partial labels.

3. Sufficient Conditions for Identifiability

We can think of each problem as requiring a specific factorization of $\tilde{\boldsymbol{P}}$ in terms of $\boldsymbol{P}$ and $\boldsymbol{\Pi}$. We say $\tilde{\boldsymbol{P}}$ is *factorizable* if there exists $(\boldsymbol{\Pi}, \boldsymbol{P}) \in \Delta_L^M \times \mathcal{P}^L$ such that $\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}\boldsymbol{P}$; we call $(\boldsymbol{\Pi}, \boldsymbol{P})$ a *factorization* of $\tilde{\boldsymbol{P}}$. Multiclass classification with label noise requires a specific ordering of the elements of $\boldsymbol{P}$; classification with partial labels requires that $\boldsymbol{\Pi}$ is consistent with $\boldsymbol{\Pi}^+$ and a specific ordering of the elements of $\boldsymbol{P}$.

A factorization is not guaranteed to exist. For example, there is no factorization in the case where $M = 3$, $L = 2$, and $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3$ are linearly independent. When a factorization exists, in general it is not unique. For instance, consider the case where $L = M$, $(\boldsymbol{\Pi}, \boldsymbol{P})$ solves (2), and $\boldsymbol{\Pi}$ is not a permutation matrix. Then, another solution is $\tilde{\boldsymbol{P}} = \boldsymbol{I}\tilde{\boldsymbol{P}}$. Furthermore, there are infinitely many solutions in the following general case.

Proposition 1 *Suppose that $\tilde{\boldsymbol{P}}$ has at least two distinct $\tilde{P}_j$s and has a factorization $(\boldsymbol{\Pi}, \boldsymbol{P})$. If there is some $\tilde{P}_i$ in the interior of* $\operatorname{conv}(P_1, \ldots, P_L)$*, then there are infinitely many distinct non-trivial factorizations of $\tilde{\boldsymbol{P}}$.*

Proof Without loss of generality, suppose that $i = 1$ and $\tilde{P}_1 \neq \tilde{P}_2$. Then, since $\tilde{P}_1$ is in the interior of $\operatorname{conv}(P_1, \ldots, P_L)$, there is some $\delta > 0$ such that for any $\alpha \in (1, 1 + \delta)$, $Q_\alpha = \alpha\tilde{P}_1 + (1-\alpha)\tilde{P}_2$ is a distribution. Then, $\operatorname{conv}(\tilde{P}_1, \ldots, \tilde{P}_L) \subseteq \operatorname{conv}(Q_\alpha, \tilde{P}_2, \ldots, \tilde{P}_L)$ and, consequently, there is some $\boldsymbol{\Pi}' \in \Delta_L^L$ such that $(\boldsymbol{\Pi}', (Q_\alpha, \tilde{P}_2, \ldots, \tilde{P}_L)^T)$ solves (2). Clearly, by varying α, there are infinitely many solutions to (2). ■

Identifiability of each problem is equivalent to the existence of a unique factorization for that problem. Therefore, to establish identifiability for the three problems, we must impose conditions on $(\boldsymbol{\Pi}, \boldsymbol{P})$ and $\boldsymbol{\Pi}^+$. To this end, we use the notion of joint irreducibility of distributions.

Definition 3 *The distributions $\{P_i\}_{1 \leqslant i \leqslant L}$ are* jointly irreducible *iff the following equivalent conditions hold*

(a) For all $I \subset [L]$ such that $1 \leqslant |I| < L$, and ϵ_i such that $\epsilon_i \geqslant 0$ and $\sum_{i \in I} \epsilon_i = \sum_{i \notin I} \epsilon_i = 1$,

$$(\sum_{i \in I} \epsilon_i P_i, \sum_{i \notin I} \epsilon_i P_i) \in IR.$$

(b) $\sum_{i=1}^L \gamma_i P_i$ is a distribution implies that $\gamma_i \geqslant 0 \, \forall i$.

Conditions *(a)* and *(b)*, whose equivalence was established by Blanchard and Scott (2014), give two ways to think about joint irreducibility. Condition *(a)* says that every convex combination of a subset of the P_is is irreducible (see Section 2.1) with respect to every convex combination of the other P_is. Condition *(b)* says that if a distribution is in the span of $P_1, \ldots, P_L$, it is in their convex hull. Joint irreducibility holds when each P_i has a region of positive probability that does not belong to the support of any of the other P_is; thus, separability (see Section 2.2) of the P_is entails joint irreducibility of $P_1, \ldots, P_L$. However, the converse is not true: the P_is can have the same support and still be jointly irreducible (e.g., P_is Gaussian with a common variance and distinct means (Scott et al., 2013)).

For all three problems, we assume that

(A) $P_1, \ldots, P_L$ are jointly irreducible.

Henceforth, unless we say otherwise, $P_1, \ldots, P_L$ are assumed to be jointly irreducible. In Appendix G, we provide experiments on real-world datasets that suggest that this assumption is reasonable.

We make different assumptions on $\mathbf{\Pi}$ for each of the three problems. For multiclass classification with label noise, we assume that

(B1) $\mathbf{\Pi}$ is invertible and $\mathbf{\Pi}^{-1}$ is a matrix with strictly positive diagonal entries and non-positive off-diagonal entries.

According to Lemma 1 below, this assumption essentially says that the problem has low noise in the sense that for each i, $\tilde{P}_i$ mostly comes from P_i. In particular, each P_i can be recovered by subtracting small multiples of $\tilde{P}_j$, $j \neq i$ from $\tilde{P}_i$. For example, consider the following case where $\mathbf{\Pi}$ satisfies **(B1)**. Suppose that there is a "common background noise" $\boldsymbol{c} \in \Delta_L$ that appears in different proportions in each of the distributions; formally, we have $\boldsymbol{\pi}_i = \gamma_i \boldsymbol{c} + (1 - \gamma_i)\boldsymbol{e}_i$ with $\gamma_i \in [0, 1)$. In other words, we shift each of the vertices $\boldsymbol{e}_i$ towards a common point $\boldsymbol{c}$ (see panel (iii) of Figure 1). See Blanchard and Scott (2014) for a proof that this setup satisfies **(B1)**. In the binary case where $M = L = 2$, **(B1)** is equivalent to the simple condition that $\pi_{1,1} + \pi_{2,2} < 1$. This assumption roughly says that in expectation the majority of labels are correct. In Section 4.3, we present Lemma 1, which gives a geometric interpretation of **(B1)**.

For the demixing problem, we assume that

(B2) $\mathbf{\Pi}$ has full column rank.

We note that **(B2)** is considerably weaker than **(B1)**, e.g., it allows $M > L$. Of course, it is natural to demand a weaker sufficient condition for demixing the mixed membership problem than muliticlass classification with label noise because the goal of the former problem is to recover any permutation of $\boldsymbol{P}$ while the goal of the latter is to recover $\boldsymbol{P}$ exactly. Nevertheless, the identifiability analysis to establish **(B2)** as a sufficient condition is also significantly more involved than the analysis of **(B1)**.

For classification with partial labels, we assume that

(B3) $\mathbf{\Pi}$ has full column rank and the columns of $\mathbf{\Pi}^+$ are unique.

The assumption that the columns of $\mathbf{\Pi}^+$ are unique says that there are no two classes that always appear together in the partial labels. In Appendix C, we argue that several of the above conditions are also necessary, or are not much stronger than what is necessary.

4. Algorithms for the Population Case

In this section, to establish that the above conditions are indeed sufficient for identifiability, we give a population case analysis of the three problems. The results on multiclass classification with label noise appeared in a conference paper (Blanchard and Scott, 2014); we refer the reader to that paper for the proofs.

4.1. Background

This paper relies on the following quantity from Blanchard et al. (2010).

Definition 4 *Given probability distributions F_0, F_1, define*

$$\kappa^*(F_0 \,|\, F_1) = \max\{\kappa \in [0,1] | \exists \text{ a distribution } G \text{ s.t. } F_0 = (1-\kappa)G + \kappa F_1\}.$$

The following Proposition from Blanchard et al. (2010) establishes some useful properties of κ^*.

Proposition 2 *Given probability distributions F_0, F_1 on a measurable space $(\mathcal{X}, \mathcal{C})$, if $F_0 \neq F_1$, then $\kappa^*(F_0 \,|\, F_1) < 1$ and the above maximum is attained for a unique distribution G (which we refer to as the* residue of F_0 wrt. F_1*). Furthermore, the following equivalent characterization holds:*

$$\kappa^*(F_0 \,|\, F_1) = \inf_{C \in \mathcal{C}, F_1(C) > 0} \frac{F_0(C)}{F_1(C)}.$$

Note that $\kappa^*(F_0 \,|\, F_1) = 0$ iff F_0 is irreducible wrt F_1. $\kappa^*(F_0 \,|\, F_1)$ can be thought of as the maximum possible proportion of F_1 in F_0. We can think of $1 - \kappa^*(F_0 \,|\, F_1)$ as a statistical distance since it is non-negative and equal to zero if and only if $F_0 = F_1$. We refer to κ^* as the two-sample κ^* operator. To obtain the residue of F_0 wrt F_1, one computes Residue$(F_0 \,|\, F_1)$ (see Algorithm 1); this is well-defined under Proposition 2 when $F_0 \neq F_1$.

In order to gain intuition about κ^*, we briefly discuss how it can be used to recover $\mathbf{\Pi}^{-1}$ in the case $L = 2$. Under conditions discussed above (Scott et al., 2013), it holds that

$$\tilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 \tilde{P}_2, \text{ and}$$
$$\tilde{P}_2 = (1-\kappa_2)P_2 + \kappa_2 \tilde{P}_1.$$

and $\kappa_1 = \kappa^*(\tilde{P}_1 \,|\, \tilde{P}_2)$ and $\kappa_2 = \kappa^*(\tilde{P}_2 \,|\, \tilde{P}_1)$. By rearranging this system of equations, we can write

$$\boldsymbol{P} = \mathbf{\Pi}^{-1}\tilde{\boldsymbol{P}} = \begin{pmatrix} \frac{1}{1-\kappa_1} & -\frac{\kappa_1}{1-\kappa_1} \\ -\frac{\kappa_2}{1-\kappa_2} & \frac{1}{1-\kappa_2} \end{pmatrix} \tilde{\boldsymbol{P}}.$$

Next, we turn to the multi-sample generalization of κ^*, which we call the multi-sample κ^* operator.

Definition 5 *Given distributions $F_0, \ldots, F_K$, define*

$$\kappa^*(F_0 \,|\, F_1, \ldots, F_K) = \max_{\boldsymbol{\mu} \in \Delta_K} \kappa^*(F_0 \,|\, \sum_{i=1}^K \mu_i F_i)$$
$$= \max\Big(\sum_{i=1}^K \nu_i : \nu_i \geqslant 0, \sum_{i=1}^K \nu_i \leqslant 1, \exists \text{ distribution } G \text{ s.t. } F_0 = (1 - \sum_{i=1}^K \nu_i)G + \sum_{i=1}^K \nu_i F_i \Big). \quad (3)$$

Algorithm 1 Residue($F_0 \,|\, F_1$)

1: $\kappa \longleftarrow \kappa^*(F_0 \,|\, F_1)$
2: **return** $\frac{F_0 - \kappa F_1}{1-\kappa}$

Algorithm 2 MultiResidue($F_0 \,|\, \{F_1, \ldots, F_K\}$)

1: $(\nu_1, \ldots, \nu_K)^T \longleftarrow (\nu_1', \ldots, \nu_K')^T$ achieving the maximum in $\kappa^*(F_0 \,|\, F_1, \ldots, F_K)$
2: **return** $\frac{F_0 - \sum_{i=1}^K \nu_i F_i}{1 - \sum_{i=1}^K \nu_i}$

Blanchard and Scott (2014) establish the equivalence in line (3), as well as Lemma 15, which shows that the outer maximum is always attained at some $\boldsymbol{\mu} \in \Delta_K$, i.e., κ^* is well-defined. Although there is always a G achieving the max, it is not necessarily unique. Any G attaining the maximum is called a *maximizer* of $\kappa^*(F_0 \,|\, F_1, \ldots, F_K)$. The algorithm MultiResidue($F_0 \,|\, \{F_1, \ldots, F_K\}$) returns one of these G (see Algorithm 2). If G is unique, we call G the *multi-sample residue of* F_0 *with respect to* $\{F_1, \ldots, F_K\}$. Under our proposed sufficient conditions, certain residues are shown to exist, and our decontamination methods compute such residues via Algorithm 2. In Section 4.3, we discuss Lemma 1, which establishes useful conditions under which a multi-sample residue exists and is equal to one of the vertices of Δ_L.

In general, one cannot express the multi-sample version of κ^* in terms of the two-sample version. However, it is possible in some special cases. For example, if one had access to feasible $\nu_1, \ldots, \nu_K$ that attain the optimum in (3), then it holds that $\kappa^*(F_0 \,|\, F_1, \ldots, F_K) = \kappa^*(F_0 \,|\, \frac{\sum_{i=1}^K \nu_i F_i}{\sum_{i=1}^K \nu_i})$. Further, it is possible to replace the multi-sample κ^* with several calls of the two-sample κ^* when $K = L - 1$, $F_i = P_i$ for all $i \neq 0$ and $F_0 = \sum_{i=1}^L \alpha_i P_i$ where $\sum_i \alpha_i = 1$ and $\forall i \ \alpha_i > 0$ (see Lemmas 6 and 13).

We remark that in previous work that assumes P_i are probability vectors, distributions are compared using l_p distances. By contrast, in our setting of general probability spaces, we use κ^* to compare different distributions.

4.2. Mixture Proportions

Recall that we assume that $P_1, \ldots, P_L$ are jointly irreducible. If $\boldsymbol{\eta} \in \mathbb{R}^L$ and $Q = \boldsymbol{\eta}^T \boldsymbol{P}$, we say that $\boldsymbol{\eta}$ is the *mixture proportion* of Q. Since by Lemma 16, joint irreducibility of $P_1, \ldots, P_L$ implies linear independence of $P_1, \ldots, P_L$, mixture proportions are well-defined, i.e., the mixture proportions are unique.

An important feature of our decontamination strategy is recovering various mixture proportions in the simplex Δ_L. To make this precise, we introduce the following definitions. If $i \in [L]$, we say that $\mathrm{conv}(\{\boldsymbol{e}_j : j \neq i\})$ is a *face* of the simplex Δ_L; if $A \subset [L]$ and $|A| = k$, we also say that $\mathrm{conv}(\{\boldsymbol{e}_j : j \in A\})$ is a *k-face* of Δ_L. If $\boldsymbol{\eta} \in \mathbb{R}^L$, Q is a distribution, and $Q = \boldsymbol{\eta}^T \boldsymbol{P}$, we say that $\mathcal{N}(Q) = \mathcal{N}(\boldsymbol{\eta}) = \{j : \eta_j > 0\}$ is the *support set* of $\boldsymbol{\eta}$ or the support set of Q. Note that by joint irreducibity, $\mathcal{N}(\boldsymbol{\eta})$ consists of the indices of all the nonzero entries in the mixture proportion $\boldsymbol{\eta}$. Finally, for $\boldsymbol{\eta}_i \in \Delta_L$, and $Q_i = \boldsymbol{\eta}_i^T \boldsymbol{P}$ for $i = 1, 2$, we

Algorithm 3 Multiclass($\tilde{P}_1, \ldots, \tilde{P}_L$)

for $i = 1, \ldots, L$ **do**
  $Q_i \longleftarrow$ MultiResidue($\tilde{P}_i \,|\, \{\tilde{P}_j : j \neq i\}$)
end for
return $(Q_1, \ldots, Q_L)^T$

say that the distributions Q_1 and Q_2 (or the mixture proportions $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$) are *on the same face* of the simplex Δ_L if there exists $j \in [L]$ such that $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \text{conv}(\{\boldsymbol{e}_k : k \neq j\})$.

The heart of our approach is that under joint irreducibility, one can interchange distributions $Q_1, \ldots, Q_K$ and their mixture proportions $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$, as indicated by the following Proposition. We note that it is valid to to apply the κ^* operator to $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ since they can be viewed as discrete probability distributions over $[L]$.

Proposition 3 *Let $Q_i = \boldsymbol{\eta}_i^T \boldsymbol{P}$ for $i \in [L]$ and $\boldsymbol{\eta}_i \in \Delta_L$. Suppose $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L$ are linearly independent and $P_1, \ldots, P_L$ are jointly irreducible. Then,*

1. *for any $i \in [L]$ and $A \subseteq [L] \backslash \{i\}$, $\kappa^*(Q_i \,|\, \{Q_j : j \in A\}) = \kappa^*(\boldsymbol{\eta}_i \,|\, \{\boldsymbol{\eta}_j : j \in A\}) < 1$,*
2. *for any $i \in [L]$ and $A \subseteq [L] \backslash \{i\}$, a maximizer of $\kappa^*(Q_i \,|\, \{Q_j : j \in A\})$ exists, and*
3. *$\boldsymbol{\gamma} \in \Delta_L$ is a maximizer to $\kappa^*(\boldsymbol{\eta}_i \,|\, \{\boldsymbol{\eta}_j : j \in A\})$ if and only if $G = \boldsymbol{\gamma}^T \boldsymbol{P}$ is a maximizer to $\kappa^*(Q_i \,|\, \{Q_j : j \in A\})$.*

In words, this proposition says that the optimization problem given by $\kappa^*(Q_i \,|\, \{Q_j : j \in A\})$ is equivalent to the optimization problem given by $\kappa^*(\boldsymbol{\eta}_i \,|\, \{\boldsymbol{\eta}_j : j \in A\})$. *Thus, joint irreducibility of $P_1, \ldots, P_L$ and linear independence of the mixture proportions enable a reduction of each of the three problems to a geometric problem where the goal is to recover the vertices of a simplex by applying κ^* to points (i.e., the mixture proportions) in the simplex.* This makes the figures below valid for general distributions (see Figures 1, 2, 3, and 4).

4.3. Multiclass Classification with Label Noise

Our algorithm for multiclass classification with label noise is by far the simplest of the three. It simply computes a maximizer of $\kappa^*(\tilde{P}_i \,|\, \{\tilde{P}_j : j \neq i\})$ for every $i \in [L]$.

Theorem 1 *Let $P_1, \ldots, P_L$ be jointly irreducible and $\boldsymbol{\Pi}$ satisfy **(B1)**. Then, Multiclass($\tilde{P}_1, \ldots, \tilde{P}_L$) returns $\boldsymbol{Q} \in \mathcal{P}^L$ such that $\boldsymbol{Q} = \boldsymbol{P}$.*

The proof of this result has two main ideas. First, it applies the one-to-one correspondence established in Proposition 3 between the maximizers of $\kappa^*(\tilde{P}_i \,|\, \{\tilde{P}_j : j \neq i\})$ and the maximizers of $\kappa^*(\boldsymbol{\pi}_i \,|\, \{\boldsymbol{\pi}_j : j \neq i\})$.

Second, the proof shows that $\kappa^*(\boldsymbol{\pi}_i \,|\, \{\boldsymbol{\pi}_j : j \neq i\})$ is well-behaved in the sense that the residue of $\boldsymbol{\pi}_i$ wrt $\{\boldsymbol{\pi}_j : j \neq i\}$ is $\boldsymbol{e}_i$. The key idea is encapsulated in the following Lemma from Blanchard and Scott (2014).

Lemma 1 (Blanchard and Scott, 2014) *The following conditions on $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L$ are equivalent:*

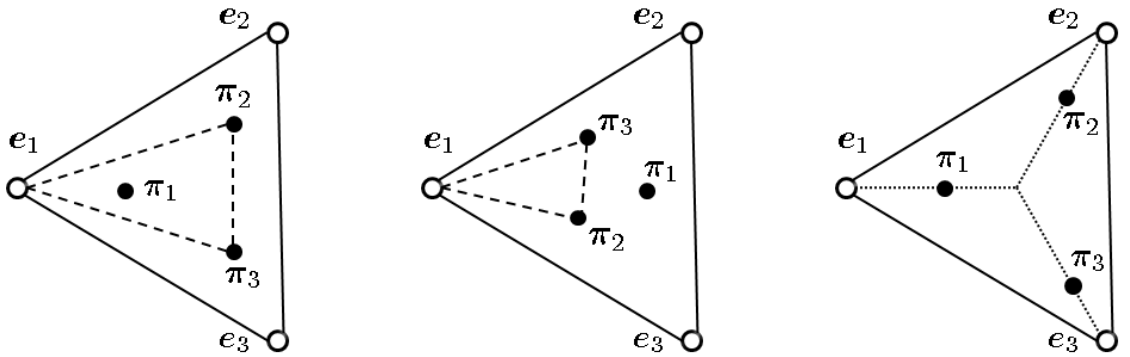

Figure 1: Illustration of the **(B1)** when there are $L = 3$ classes where $\boldsymbol{e}_i$ denotes the ith unit vector. Panel (i): Low noise, $\boldsymbol{\Pi}$ recoverable. Each $\boldsymbol{\pi}_l$ can be written as a convex combination of $\boldsymbol{e}_l$ and the other two $\boldsymbol{\pi}_j$ (with a positive weight on $\boldsymbol{e}_l$), depicted here for $l = 1$. Panel (ii): High noise, $\boldsymbol{\Pi}$ not recoverable. Panel (iii): The setting of "common background noise" described in the text.

1. *For each i, the residue of $\boldsymbol{\pi}_i$ with respect to $\{\boldsymbol{\pi}_j, j \neq i\}$ is $\boldsymbol{e}_i$.*

2. *For every i there exists a decomposition $\boldsymbol{\pi}_i = \kappa_i \boldsymbol{e}_i + (1 - \kappa_i)\boldsymbol{\pi}_i'$ where $\kappa_i > 0$ and $\boldsymbol{\pi}_i'$ is a convex combination of $\boldsymbol{\pi}_j$ for $j \neq i$.*

3. *$\boldsymbol{\Pi}$ is invertible and $\boldsymbol{\Pi}^{-1}$ is a matrix with strictly positive diagonal entries and non-positive off-diagonal entries.*

This lemma establishes that under **(B1)**, for each i, the residue of $\boldsymbol{\pi}_i$ with respect to $\{\boldsymbol{\pi}_j, j \neq i\}$ is $\boldsymbol{e}_i$. The main step in the proof of this Lemma is establishing that *3* implies *1*. The argument identifies the residue of $\boldsymbol{\pi}_i$ with respect to $\{\boldsymbol{\pi}_j\}_{j \neq i}$ by reformulating the linear program in $\kappa^*(\boldsymbol{\pi}_i \,|\, \{\boldsymbol{\pi}_j\}_{j \neq i})$ such that the objective is to maximize $\boldsymbol{e}_i^t \boldsymbol{\Pi}^{-1} \gamma$ subject to some appropriately defined constraint. By the structure of $\boldsymbol{\Pi}^{-1}$ assumed in **(B1)**, it follows that the $\gamma \in \Delta_L$ that maximizes this objective is $\boldsymbol{e}_i$, and it can further be shown that this maximizer satisfies the other constraints.

Thus, combining the above two ideas yields the result. In addition, Lemma 1 provides geometric intuition as to when **(B1)** is satisfied through condition *2*. Figure 1 illustrates the case $L = 3$. See Panel (i) for an example where condition (b) is satisfied and Panel (ii) for an example where (b) is not satisfied.

4.4. Demixing Mixed Membership Models

In this section, we assume that $M = L$; we consider the nonsquare case in the appendix. For certain simple cases of mixture proportions, a straightforward resampling strategy can be used to reduce the problem of demixing mixed membership models to multiclass classi-

fication with label noise. For example, suppose that there are $L = 3$ classes and

$$\mathbf{\Pi} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{4}$$

Inspection shows that the inverse of $\mathbf{\Pi}$ does not satisfy the condition in **(B1)** and, therefore, one cannot simply apply Algorithm 3. A simple procedure to circumvent this issue is to resample from the contaminated distributions to obtain the following distributions:

$$\tilde{Q}_1 = \frac{1}{2}\tilde{P}_1 + \frac{1}{2}\tilde{P}_2,\ \tilde{Q}_2 = \frac{1}{2}\tilde{P}_1 + \frac{1}{2}\tilde{P}_3,\ \text{and}\ \tilde{Q}_3 = \frac{1}{2}\tilde{P}_2 + \frac{1}{2}\tilde{P}_3.$$

Then, it can be shown that the resulting mixing matrix

$$\tilde{\mathbf{\Pi}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

associated with the $\tilde{Q}_i$s satisfies the conditions of Lemma 1 so that Multiclass$(\tilde{Q}_1, \tilde{Q}_2, \tilde{Q}_3)$ gives the desired solution. However, this approach breaks down for most possible mixing matrices. Thus, the challenge is to develop an algorithm that works for a large class of mixture proportions and does not rely on knowledge of the mixture proportions. To meet this challenge, we propose the Demix algorithm.

The Demix algorithm is recursive. Let $S_1, \ldots, S_K$ denote K contaminated distributions. In the base case, the algorithm takes as its input two contaminated distributions S_1 and S_2. It returns Residue$(S_1 \,|\, S_2)$ and Residue$(S_2 \,|\, S_1)$, which are a permutation of the two base distributions (see Figure 2). When $K > 2$, Demix uses a subroutine FindFace (see Algorithm 5) to find $K - 1$ distributions $R_2, \ldots, R_K$ on the same $(K-1)$-face. FindFace iteratively generates candidates for distributions on the same $(K-1)$-face, which it tests using FaceTest (see Algorithm 6). FaceTest$(S_1, \ldots, S_{K-1})$ determines whether a set of distributions $S_1, \ldots, S_{K-1}$ are on the interior of the same face by using the two-sample κ^* operator; equivalently, it tests whether there exists a pair of distributions S_i and S_j such that S_i is irreducible with respect to S_j. Once Demix finds $K-1$ distributions $R_2, \ldots, R_K$ on the same $(K-1)$-face, it recursively applies Demix to $R_2, \ldots, R_K$ to obtain distributions $Q_1, \ldots, Q_{K-1}$ that are a permutation of $K-1$ of the base distributions. Subsequently, the algorithm computes a maximizer Q_K of $\kappa^*(\frac{1}{K}\sum_{i=1}^{K} S_i \,|\, Q_1, \ldots, Q_{K-1})$. Since $Q_1, \ldots, Q_{K-1}$ are a permutation of $K-1$ of the base distributions, the maximizer Q_K is guaranteed to be unique and to be the remaining base distribution (see Figure 3 for an execution of the algorithm).

A number of remarks are in order regarding the Demix algorithm. First, although we compute the residue of $\frac{1}{n}S_i + \frac{n-1}{n}Q$ wrt S_1 for each $i \neq 1$, there is nothing special about the distribution S_1. We could replace S_1 with any S_j where $j \in [K]$, provided that we adjust the rest of the algorithm accordingly. Second, we can replace the sequence $\{\frac{n-1}{n}\}_{n=1}^{\infty}$ with any sequence $\alpha_n \nearrow 1$. Finally, we could replace line 7 with the following sequence of steps: for $i = 1, \ldots, K-1$, compute $Q_K \longleftarrow$ Residue$(Q_K \,|\, Q_i)$ (see Lemma 13). Then, the algorithm would only use the two-sample κ^* operator. We use such an algorithm in the finite-sample setting.

Algorithm 4 Demix$(S_1, \ldots, S_K)$

Input: $S_1, \ldots, S_K$ are distributions

if $K = 2$ **then**
  return $(\text{Residue}(S_1 \mid S_2), \text{Residue}(S_2 \mid S_1))^T$
else
  $(R_2, \ldots, R_K)^T \longleftarrow \text{FindFace}(S_1, \ldots, S_K)$
  $(Q_1, \ldots, Q_{K-1})^T \longleftarrow \text{Demix}(R_2, \ldots, R_K)$
  $Q_K \longleftarrow \frac{1}{K}\sum_{i=1}^K S_i$
  $Q_K \longleftarrow \text{MultiResidue}(Q_K \mid Q_1, \ldots, Q_{K-1})$
  return $(Q_1, \ldots, Q_K)^T$
end if

Algorithm 5 FindFace$(S_1, \ldots, S_K)$

Input: $S_1, \ldots, S_K$ are distributions

$Q \longleftarrow$ uniformly distributed element in $\text{conv}(S_2, \ldots, S_K)$
for $n = 1, 2, \ldots$ **do**
  Set $R_i \longleftarrow \text{Residue}(\frac{1}{n}S_i + \frac{n-1}{n}Q \mid S_1)$ for all $i \in \{2, \ldots, K\}$
  if $\text{FaceTest}(R_2, \ldots, R_K)$ **then**
    return $(R_2, \ldots, R_K)^T$
  end if
end for

We also remark that a simplified version of Demix solves the demixing mixed membership models problem if we assume **(B1)** from the multiclass label noise setting. In that case, finding $L-1$ distributions on the same face can be accomplished by simply computing $Q_i \longleftarrow \text{MultiResidue}(\tilde{P}_i \mid \{\tilde{P}_j\}_{j \neq i})$ for $i = 2, \ldots, L$. Indeed, then, each Q_i is equal to P_i and P_1 can be obtained by computing $\text{MultiResidue}(\tilde{P}_1 \mid \{Q_j\}_{j=2,\ldots,L})$.

We establish the following theorem.

Theorem 2 *Let $P_1, \ldots, P_L$ be jointly irreducible and $\mathbf{\Pi}$ have full column rank. Then, with probability 1, Demix$(\tilde{\boldsymbol{P}})$ returns a permutation of $\boldsymbol{P}$.*

We briefly sketch three key aspects of the proof. First, in the FindFace subroutine, sampling Q uniformly at random from $\text{conv}(S_2, \ldots, S_K)$ ensures that w.p. 1 $\text{Residue}(Q \mid S_1)$ is on the interior of a face of the simplex. Then, conditional on this event, we show that by a continuity property of $\text{Residue}(\cdot \mid S_1)$ there is a large enough n such that $R_2, \ldots, R_K$ are on the same face of the simplex Δ_{K-1} (see panels (c) and (d) of Figure 3). Second, Proposition 8 in Appendix D establishes that the subroutine $\text{FaceTest}(R_2, \ldots, R_K)$ returns 1 if and only if $R_2, \ldots, R_K$ are on the same face of the simplex. Combining the above two observations implies that eventually $\text{FindFace}(S_1, \ldots, S_K)$ terminates at which point $\{R_k\}_{k \in [K]\setminus\{1\}} \subset \{P_k\}_{k \in [K]\setminus\{l\}}$ for some $l \in [K]$. The final key observation is that $\{R_k\}_{k \in [K]\setminus\{1\}}$ and $\{P_k\}_{k \in [K]\setminus\{l\}}$ form an instance of the demixing problem that satisfies the sufficient conditions **(A)** and **(B2)** (see Figure 2). Therefore, this instance can be solved recursively.

Algorithm 6 FaceTest$(S_1, \ldots, S_K)$

1: Set $\boldsymbol{Z}_{i,j} \coloneqq \mathbf{1}\{\kappa^*(S_i \,|\, S_j) > 0\}$ for all i and j
2: **if** $\boldsymbol{Z}$ has a zero off-diagonal entry **then**
3: **return** 0
4: **else**
5: **return** 1
6: **end if**

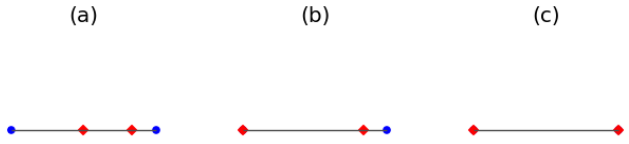

Figure 2: In (a), we consider a demixing problem where there are two classes and $M = L$ (the base case of Algorithm 4). The diamonds represent the mixture proportions of $\tilde{P}_1$ and $\tilde{P}_2$. The circles represent the base distributions. In (b), the residue of a contaminated distribution wrt the other contaminated distribution is computed (line 3), yielding a base distribution. In (c), the residue is computed again switching the roles of the contaminated distributions (line 4); this yields the remaining base distribution.

4.5. Classification with Partial Labels

As in the case of demixing mixed membership models, a simple resampling strategy works in certain nice settings of classification with partial labels. For example, consider an instance of classification with partial labels with the mixing matrix from equation (4). The resampling procedure that yields $\tilde{Q}_1, \tilde{Q}_2, \tilde{Q}_3$ (described in Section 4.4) also works here. Nevertheless, as in demixing mixed membership models, this approach does not meet our goal of an algorithm that solves a broad class of mixing matrices and partial labels.

Indeed, we observe that the partial labels do not provide enough information for choosing the resampling weights. Consider another instance of the problem with the same partial label matrix:

$$\mathbf{\Pi} = \begin{pmatrix} \frac{1}{10} & \frac{9}{10} & 0 \\ \frac{9}{10} & 0 & \frac{1}{10} \\ 0 & \frac{1}{10} & \frac{9}{10} \end{pmatrix}. \tag{5}$$

Applying the resampling approach to (5) can be shown to fail by observing that the inverse of the resampled mixing matrix does not satisfy condition *3* of Lemma 1. Thus, although the

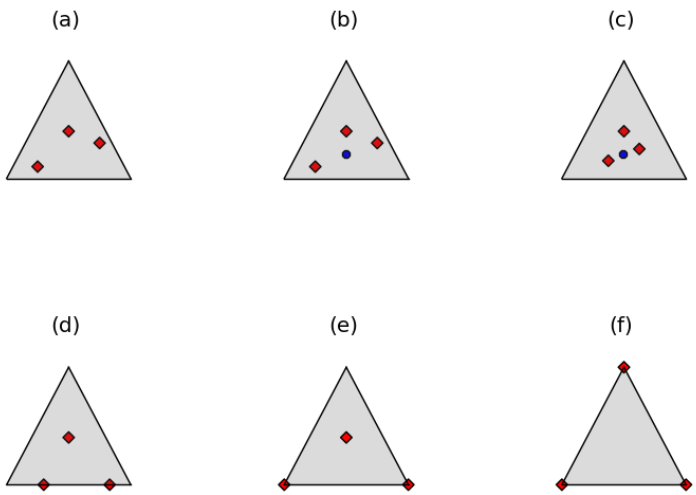

Figure 3: In (a), we consider a demixing problem where there are three classes and $M = L$. The diamonds represent the mixture proportions of $\tilde{P}_1, \tilde{P}_2$ and $\tilde{P}_3$. In (b), the blue circle is a random distribution chosen in the convex hull of two of the distributions (line 7). In (c), two of the distributions are resampled so that their residues wrt the other distribution are on the same face of the simplex (lines 12-15). In (d), these particular residues are computed (lines 12-15). In (e), two of the distributions are demixed (lines 3-5). In (f), the residue of the final distribution wrt the final two demixed distributions is computed to obtain the final demixing (line 18-21).

Algorithm 7 PartialLabel($\mathbf{\Pi}^+, (\tilde{P}_1, \ldots, \tilde{P}_M)^T$)

1: **for** $i = 1, \ldots, L$ **do**
2: $\quad Q_i \longleftarrow$ uniformly random distribution in $\text{conv}(\tilde{P}_1, \ldots, \tilde{P}_M)$
3: **end for**
4: **for** $k = 2, 3, \ldots$ **do**
5: $\quad (W_1, \ldots, W_L)^T \longleftarrow$ GenerateCandidates$(k, (Q_1, \ldots, Q_L)^T)$
6: $\quad$ (FoundVertices, $\boldsymbol{C}$) $\longleftarrow$ VertexTest$(\mathbf{\Pi}^+, \tilde{P}_1, \ldots, \tilde{P}_M, W_1, \ldots, W_L)$
7: $\quad$ **if** FoundVertices **then**
8: $\quad\quad$ **return** $\boldsymbol{C}(W_1, \ldots, W_L)^T$
9: $\quad$ **end if**
10: **end for**

problem instances (4) and (5) have the same partial label matrix, the resampling procedure only works for one of these.

Next, we turn to presenting an algorithm that solves classification with partial labels for a wide class of mixing matrices and partial labels. We propose the PartialLabel algorithm

Algorithm 8 GenerateCandidates$(k, (Q_1, \ldots, Q_L)^T)$

Set $W_i \longleftarrow Q_i$ for all $i \in [K]$
for $i = 1, \ldots, L$ **do**
  $\bar{Q}_i \longleftarrow \frac{1}{L-1}[\sum_{j>i} Q_j + \sum_{j<i} W_j]$
  $W_i \longleftarrow \text{MultiResidue}(\frac{1}{k}Q_i + (1 - \frac{1}{k})\bar{Q}_i \,|\, \{Q_j\}_{j>i} \cup \{W_j\}_{j<i})$
end for
return $(W_1, \ldots, W_L)^T$

Algorithm 9 VertexTest$(\mathbf{\Pi}^+, (\tilde{P}_1, \ldots, \tilde{P}_M)^T, (Q_1, \ldots, Q_L)^T)$

Form the matrix $Y_{i,j} \coloneqq \mathbf{1}\{\kappa^*(Q_i \,|\, Q_j) > 0\}$
if $\boldsymbol{Y}$ has a non-zero off-diagonal entry **then**
  return $(0, \mathbf{0})$
end if
Form the matrix $Z_{i,j} \coloneqq \mathbf{1}\{\kappa^*(\tilde{P}_i \,|\, Q_j) > 0\}$
Use any algorithm that finds a permutation matrix $\boldsymbol{C}$ such that $\boldsymbol{ZC} = \mathbf{\Pi}^+$ (if it exists)
if such a permutation matrix $\boldsymbol{C}$ exists **then**
  return $(1, \boldsymbol{C}^T)$
else
  return $(0, \mathbf{0})$
end if

(see Algorithm 7). PartialLabel proceeds by iteratively creating sets of candidate distributions $\boldsymbol{W} \coloneqq (W_1, \ldots, W_L)^T$ via the subroutine CreateCandidates (see Algorithm 8). Given each $\boldsymbol{W}$, it runs an algorithm VertexTest (see Algorithm 9) that uses $\tilde{\boldsymbol{P}}$ and the partial label matrix $\mathbf{\Pi}^+$ to determine whether $\boldsymbol{W}$ is a permutation of the base distributions $\boldsymbol{P}$. If $\boldsymbol{W}$ is a permutation of $\boldsymbol{P}$, VertexTest constructs the corresponding permutation matrix for relating these distributions. If not, it reports failure and the PartialLabel algorithm increments k and finds another set of candidate distributions.

The VertexTest algorithm proceeds as follows on a vector of candidate distributions $\boldsymbol{Q} \coloneqq (Q_1, \ldots, Q_L)^T$. First, it determines whether there are two distinct distributions Q_i, Q_j such that Q_i is not irreducible wrt Q_j, in which case $\boldsymbol{Q}$ cannot be a permutation of $\boldsymbol{P}$. If there is such a pair, it reports failure. Otherwise, it forms the matrix $Z_{i,j} \coloneqq \mathbf{1}\{\kappa^*(\tilde{P}_i \,|\, Q_j) > 0\}$ and uses any algorithm that finds a permutation $\boldsymbol{C}$ (if it exists) of the columns of $\boldsymbol{Z}$ to match the columns of $\mathbf{\Pi}^+$. If such a permutation $\boldsymbol{C}$ exists, it returns $\boldsymbol{C}^T$ and, as we show in Lemma 7, $\boldsymbol{C}^T\boldsymbol{Q} = \boldsymbol{P}$; otherwise, VertexTest reports failure.

We remark that finding the permutation in line 6 of Algorithm 9 is not NP-hard. One algorithm (but most likely not the most efficient) proceeds as follows: define a total ordering on the columns of binary matrices. Sort the columns of $\boldsymbol{Z}$ and $\mathbf{\Pi}^+$ according to this total ordering. Check whether the resulting matrices are equal.

The following theorem gives our identification result for classification with partial labels.

Theorem 3 *Suppose that $P_1, \ldots, P_L$ are jointly irreducible, $\mathbf{\Pi}$ has full column rank, and the columns of $\mathbf{\Pi}^+$ are unique. Then, PartialLabel($\mathbf{\Pi}^+, (\tilde{\boldsymbol{P}})^T$) returns $\boldsymbol{R} \in \mathcal{P}^L$ such that $\boldsymbol{R} = \boldsymbol{P}$.*

There are two key ideas to the proof of Theorem 3. First, the randomization in line 2 of Algorithm 7 ensures through a linear independence argument that with probability 1, the operation MultiResidue($\frac{1}{k}Q_i + (1-\frac{1}{k})\bar{Q}_i \mid \{Q_j\}_{j>i} \cup \{W_j\}_{j<i}$) in line 4 of Algorithm 8 is well-defined. Second, in the GenerateCandidates algorithm, let $Q_j = \boldsymbol{\tau}_j^T \boldsymbol{P}$ and $W_j = \boldsymbol{\gamma}_j^T P$. We make the simple observation that the affine hyperplane given by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{i-1}, \boldsymbol{\tau}_{i+1}, \ldots, \boldsymbol{\tau}_L$ bisects Δ_L such that $\boldsymbol{\tau}_i$ and a nonempty subset of $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_L\} \backslash \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{i-1}\}$ are in the same halfspace. Using this observation, we show that for large enough k, W_i is one of the base distributions and is distinct from all W_j with $j < i$.

The VertexTest algorithm connects the demixing problem and classification with partial labels by showing that any algorithm that solves the demixing problem can be used as a subroutine to solve classification with partial labels. For example, consider the following algorithm for classification with partial labels. First, use the Demix algorithm to obtain a permutation $\boldsymbol{Q}$ of the base distributions $\boldsymbol{P}$. Second, use VertexTest to find the permutation matrix relating $\boldsymbol{Q}$ and $\boldsymbol{P}$. This alternate algorithm is the basis of our finite sample algorithm for classification with partial labels (see Section 5.3 for a more thorough discussion).

5. Estimators for the Finite Sample Setting

In this section, we develop the estimation theory to treat the three problems in the finite sample setting. Let $\mathcal{X} = \mathbb{R}^d$ be equipped with the standard Borel σ-algebra $\mathcal{C}$ and $\tilde{P}_1, \ldots, \tilde{P}_L$ be probability distributions on this space. Suppose that we observe for $i = 1, \ldots, L$,

$$X_1^i, \ldots, X_{n_i}^i \overset{i.i.d.}{\sim} \tilde{P}_i.$$

Let $\mathcal{E}$ be any Vapnik-Chervonenkis (VC) class with VC-dimension $V < \infty$, containing the set of all open balls, all open rectangles, or some other collection of sets that generates the Borel σ-algebra $\mathcal{C}$. For example, $\mathcal{E}$ could be the set of all open balls wrt the Euclidean distance, in which case $V = d + 1$. Define $\epsilon_i(\delta_i) \equiv 3\sqrt{\frac{V \log(n_i+1) - \log(\delta_i/2)}{n_i}}$ for $i = 1, \ldots, L$. Our estimators are based on the VC inequality (Devroye et al., 1996). This inequality says that for each $i \in [L]$, and $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\sup_{E \in \mathcal{E}} |\tilde{P}_i(E) - \tilde{P}_i^\dagger(E)| \leqslant \epsilon_i(\delta)$$

where the *empirical distribution* is given by $\tilde{P}_i^\dagger(E) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\{X_j^i \in E\}}$.

5.1. Multiclass Classification with Label Noise

Let $F_0^\dagger, \ldots, F_M^\dagger$ denote the empirical distributions based on i.i.d. random samples from respective distributions $F_0, \ldots, F_M$. We introduce the following estimator of the multi-sample κ^*:

$$\widehat{\kappa}(F_0^\dagger \mid F_1^\dagger, \ldots, F_M^\dagger) = \max_{\boldsymbol{\mu} \in \Delta_M} \inf_{E \in \mathcal{E}} \frac{F_0^\dagger(E) + \epsilon_0(\frac{1}{n_0})}{(\sum_{i=1}^M \mu_i F_i^\dagger(E) - \sum_i \mu_i \epsilon_i(\frac{1}{n_i}))_+} \tag{6}$$

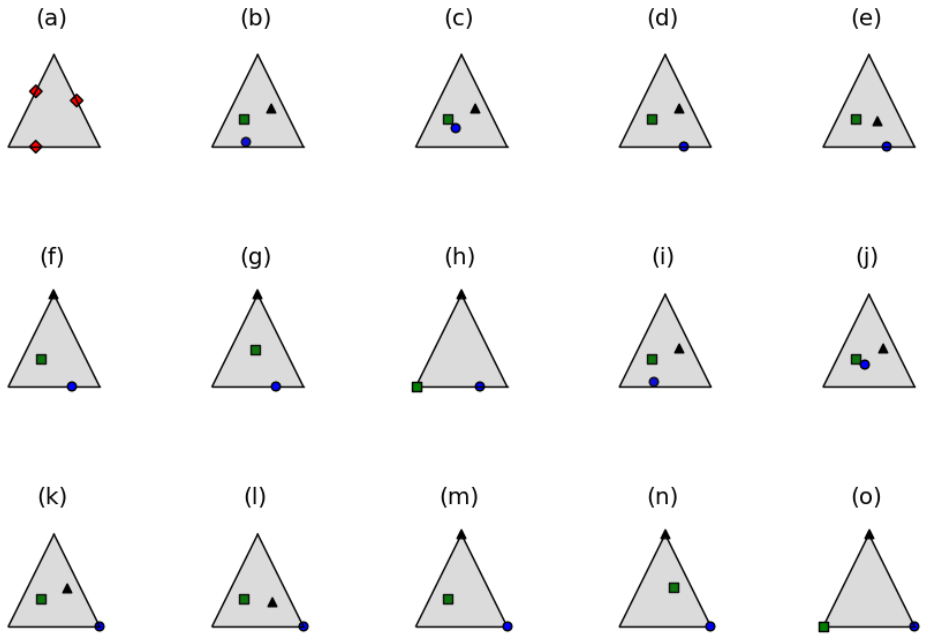

Figure 4: (a) depicts an instance of the partial label problem where there are $L = 3$ classes, $M = 3$ partial labels, and each partial label only contains two of the classes. In (a), the red diamonds represent the mixture proportions of the distributions $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3$. In (b), three distributions Q_1, Q_2, Q_3 are sampled uniformly randomly from the convex hull of $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3$; the blue circle, black triangle, and green square represent their mixture proportions. Figures (c)-(h) show how the algorithm generates a set of candidate distributions $(W_1^{(2)}, W_2^{(2)}, W_3^{(2)})^T$ with $k = 2$. In (h), PartialLabel runs VertexTest on $(W_1^{(2)}, W_2^{(2)}, W_3^{(2)})^T$ and determines that $(W_1^{(2)}, W_2^{(2)}, W_3^{(2)})^T$ is not a permutation of $(P_1, P_2, P_3)^T$. In (i)-(o), PartialLabel begins again with Q_1, Q_2, Q_3 and executes the same series of steps with $k = 3$, generating $(W_1^{(3)}, W_2^{(3)}, W_3^{(3)})^T$. In (o), it runs VertexTest on $(W_1^{(3)}, W_2^{(3)}, W_3^{(3)})^T$ and determines that $(W_1^{(3)}, W_2^{(3)}, W_3^{(3)})^T$ is a permutation of $(P_1, P_2, P_3)^T$.

where the ratio is defined to be ∞ when the denominator is zero. This estimator arises from applying the VC inequality to the following expression:

$$\kappa^*(F_0 \,|\, F_1, \ldots, F_M) = \max_{\boldsymbol{\mu} \in \Delta_M} \kappa^*(F_0 \,|\, \sum_{i=1}^{M} \mu_i F_i) = \max_{\boldsymbol{\mu} \in \Delta_M} \inf_{E \in \mathcal{E}, \sum_{i=1}^{M} \mu_i F_i(E) > 0} \frac{F_0(E)}{\sum_{i=1}^{M} \mu_i F_i(E)},$$

where the last equality uses Proposition 2. Let $\widehat{\boldsymbol{\mu}}$ denote a point where the maximum is achieved in (6). Then, $\widehat{\boldsymbol{\nu}} \coloneqq \widehat{\kappa}\widehat{\boldsymbol{\mu}}$ estimates the vector $(\nu_1, \ldots, \nu_M)$ attaining the maximum

in (3). See Proposition 2 of Blanchard and Scott (2014) to find a proof that the proposed estimator is consistent.

Based on this estimator, we introduce estimators that under the assumptions of Theorem 1 converge to the base distributions uniformly in probability. Let $\boldsymbol{n} \equiv (n_1, \ldots, n_L)$; we write $\boldsymbol{n} \longrightarrow \infty$ to indicate that $\min_i n_i \longrightarrow \infty$.

Theorem 4 *Let $(\widehat{\nu}_{i,j})_{j \neq i}$ be a vector attaining the maximum in the definition of $\widehat{\kappa}_i \coloneqq \widehat{\kappa}(\tilde{P}_i^\dagger \,|\, \{\tilde{P}_j^\dagger : j \neq i\})$ and*

$$\widehat{Q}_i = \frac{\tilde{P}_i^\dagger - \sum_{j \neq i} \widehat{\nu}_{i,j} \tilde{P}_j^\dagger}{1 - \widehat{\kappa}_i}.$$

Then, under the assumptions of Theorem 1, $\forall i = 1, \ldots, L$, $\sup_{E \in \mathcal{E}} |\widehat{Q}_i(E) - P_i(E)| \xrightarrow{i.p.} 0$ as $\boldsymbol{n} \longrightarrow \infty$.

5.2. Demixing Mixed Membership Models

In this section, we develop a novel estimator that can be used to extend the Demix algorithm to the finite sample case. Uniform convergence results typically assume access to i.i.d. samples. The challenge of developing an estimator for Demix is that because of the recursive nature of the Demix algorithm, we cannot assume access to i.i.d. samples to estimate every distribution that arises. Nonetheless, we show that uniform convergence of distributions propagates through the algorithm if we employ an estimator of κ^* with a known rate of convergence.

Let $\widehat{F}$ and $\widehat{H}$ be estimates of distributions F and H, respectively. We introduce the following estimator:

$$\widehat{\kappa}(\widehat{F} \,|\, \widehat{H}) = \inf_{E \in \mathcal{E}} \frac{\widehat{F}(E) + \gamma_{\boldsymbol{n}}}{(\widehat{H}(E) - \gamma_{\boldsymbol{n}})_+}$$

where $\gamma_{\boldsymbol{n}} = \sum_{i=1}^L \epsilon_i(\frac{1}{n_i})$. Our estimator is closely related to the estimator from Blanchard et al. (2010): if $\widehat{F}$ and $\widehat{H}$ are empirical distributions, e.g., $\widehat{F} = \tilde{P}_i^\dagger$ and $\widehat{H} = \tilde{P}_j^\dagger$, then their estimator for $\kappa^*(F \,|\, H)$ is $\inf_{E \in \mathcal{E}} \frac{\widehat{F}(E) + \epsilon_i(\frac{1}{n_i})}{(\widehat{H}(E) - \epsilon_j(\frac{1}{n_j}))_+}$. Note that our proofs only require that $\gamma_{\boldsymbol{n}}$ include the terms $\epsilon_i(\frac{1}{n_i})$ corresponding to $\tilde{P}_i$ that the estimators $\widehat{F}$ and $\widehat{H}$ use samples from; to simplify presentation, however, we include all the terms, which leads to bounds that are looser by only a constant factor.

Based on the estimator $\widehat{\kappa}$, we introduce the following estimator of the residue of F wrt H.

Definition 6 *Let $\widehat{F}$ and $\widehat{H}$ be estimators of F and H, respectively, where $F \neq H$ and let $G \longleftarrow$ Residue$(F \,|\, H)$ and $\widehat{G} \longleftarrow$ ResidueHat$(\widehat{F} \,|\, \widehat{H})$. We call $\widehat{G}$ a* ResidueHat estimator *of order $k \geqslant 1$ if (i) $F, H \in \operatorname{conv}(P_1, \ldots, P_L)$, and (ii) at least one of $\widehat{F}$ and $\widehat{H}$ is a ResidueHat estimator of order $k-1$ and the other is either an empirical distribution or a ResidueHat estimator of order less than or equal to $k-1$. We call $\widehat{G}$ a* ResidueHat estimator *of order 0 if (i) holds, and $\widehat{F}$ and $\widehat{H}$ are empirical distributions.*

Algorithm 10 ResidueHat($\widehat{F} \,|\, \widehat{H}$)

Input: $\widehat{F}, \widehat{H}$ are estimates of F, H

$\widehat{\kappa} \longleftarrow \widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$

return $\frac{\widehat{F} - \widehat{\kappa}(1-\widehat{H})}{1-\widehat{\kappa}}$

Algorithm 11 DemixHat($\widehat{S}_1, \ldots, \widehat{S}_K \,|\, \epsilon$)

Input: $\widehat{S}_1, \ldots, \widehat{S}_K$ are ResidueHat estimates

1: **if** $K = 2$ **then**
2: **return** $(\text{ResidueHat}(\widehat{S}_1 \,|\, \widehat{S}_2), \text{ResidueHat}(\widehat{S}_2 \,|\, \widehat{S}_1))^T$
3: **else**
4: $(R_2, \ldots, R_K)^T \longleftarrow \text{FindFaceHat}(\widehat{S}_1, \ldots, \widehat{S}_K \,|\, \epsilon)$
5: $(\widehat{Q}_1, \cdots, \widehat{Q}_{K-1})^T \longleftarrow \text{DemixHat}(\widehat{R}_2, \cdots, \widehat{R}_K)$
6: $\widehat{Q}_K \longleftarrow \frac{1}{K} \sum_{i=1}^{K} \widehat{S}_i$
7: **for** $i = 1, \ldots, K-1$ **do**
8: $\widehat{Q}_K \longleftarrow \text{ResidueHat}(\widehat{Q}_K \,|\, \widehat{Q}_i)$
9: **end for**
10: **return** $(\widehat{Q}_1, \cdots, \widehat{Q}_K)^T$
11: **end if**

Note that the above definition is recursive and matches the recursive structure of the Demix algorithm. We suppress the qualifier "of order k" when it is not relevant.

To use ResidueHat estimators to estimate the P_is, we build on the rate of convergence result from Scott (2015). In Scott (2015), a rate of convergence was established for an estimator of κ^* using empirical distributions; we extend these results to our setting of recursive estimators and achieve the same rate of convergence. To ensure that this rate of convergence holds for every estimate in our algorithm, we introduce the following condition.

(A″) $P_1, \ldots, P_L$ are such that $\forall i \ \text{supp}(P_i) \nsubseteq \cup_{j \neq i} \text{supp}(P_j)$.

Note that this assumption implies joint irreducibility and is a natural generalization of the separability assumption.

The following result establishes sufficient conditions under which ResidueHat estimates converge uniformly.

Proposition 4 *If $P_1, \ldots, P_L$ satisfy **(A″)** and $\widehat{G}$ is a ResidueHat estimator of a distribution $G \in \text{conv}(P_1, \ldots, P_L)$, then $\sup_{E \in \mathcal{E}} |\widehat{G}(E) - G(E)| \xrightarrow{i.p.} 0$ as $\boldsymbol{n} \longrightarrow \infty$.*

Based on the ResidueHat estimators, we introduce an empirical version of the Demix algorithm—DemixHat (see Algorithm 11). The main differences are that *(i)* we replace the Residue function with the ResidueHat function, *(ii)* we replace line 7 in the Demix algorithm with a sequence of applications of the two-sample κ^* operator, as mentioned just before Theorem 2, and *(iii)* DemixHat requires specification of a hyperparameter $\epsilon \in (0, 1)$. We replace the multi-sample κ^* with the two-sample κ^* because there is no known estimator with a rate of convergence for the multi-sample κ^*, and the rate of convergence is

Algorithm 12 FindFaceHat($\widehat{S}_1, \ldots, \widehat{S}_K \,|\, \epsilon$)

Input: $\widehat{S}_1, \ldots, \widehat{S}_K$ are ResidueHat estimates

$\widehat{Q} \longleftarrow$ uniformly distributed random element from $\text{conv}(\widehat{S}_2, \ldots, \widehat{S}_K)$
for $n = 2, 3, \ldots$ **do**
  Set $\widehat{R}_i \longleftarrow \text{ResidueHat}(\frac{1}{n}\widehat{S}_i + \frac{n-1}{n}\widehat{Q}) \,|\, \widehat{S}_1)$ for all $i \in \{2, \ldots, K\}$
  if FaceTestHat($\widehat{R}_2, \cdots, \widehat{R}_K \,|\, \epsilon$) **then**
    return $(\widehat{R}_2, \cdots, \widehat{Q}_K)^T$
  end if
end for

Algorithm 13 FaceTestHat($\widehat{Q}_1, \cdots, \widehat{Q}_K \,|\, \epsilon$)

Set $\boldsymbol{Z}_{i,j} \coloneqq \mathbf{1}\{\widehat{\kappa}(\widehat{Q}_i \,|\, \widehat{Q}_j) > \epsilon\}$ for $i \neq j$
if $\boldsymbol{Z}$ has a zero off-diagonal entry **then**
  return 0
else
  return 1
end if

essential to our consistency proof. The hyperparameter ϵ gives a tradeoff between runtime and accuracy. The runtime increases with increasing ϵ, but the amount of uncertainty about whether DemixHat executes successfully decreases with increasing ϵ.

We now state our main estimation result.

Theorem 5 *Let $\delta > 0$ and $\epsilon \in (0, 1)$. Suppose that $P_1, \ldots, P_L$ satisfy **(A″)** and $\mathbf{\Pi}$ has full rank. Then, with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, DemixHat($\tilde{P}_1^\dagger, \ldots, \tilde{P}_L^\dagger \,|\, \epsilon$) returns $(\widehat{Q}_1, \ldots, \widehat{Q}_L)$ for which there exists a permutation $\sigma : [L] \longrightarrow [L]$ such that for every $i \in [L]$,*

$$\sup_{E \in \mathcal{E}} |\widehat{Q}_i(E) - P_{\sigma(i)}(E)| < \delta.$$

5.3. Classification with Partial Labels

In this section, we present a finite sample algorithm for the decontamination of a partial label model (see Algorithm 14). This algorithm is based on a different approach from PartialLabel (Algorithm 7): it combines DemixHat with an empirical version of the VertexTest algorithm (see Algorithm 17). The reason for this is that we have an estimator with a rate of convergence for the two-sample κ^*, whereas there is no known estimator with a rate of convergence for the multi-sample κ^*. We leverage this rate of convergence to prove the consistency of our algorithm.

We make an assumption that simplifies our algorithm: $\mathbf{\Pi}^+$ satisfies

(D) there does not exist $i, j \in [L]$ such that $\mathbf{\Pi}^+_{i,:} = \boldsymbol{e}_j^T$.

In words, this says that there is no contaminated distribution $\tilde{P}_i$ and base distribution P_j such that $\tilde{P}_i = P_j$. We emphasize that we make this assumption only to simplify the

Algorithm 14 PartialLabelHat($\mathbf{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T \,|\, \epsilon$)

1: $(\widehat{Q}_1, \ldots, \widehat{Q}_L)^T \longleftarrow$ DemixHat($\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger \,|\, \epsilon$)
2: (FoundVertices, $\boldsymbol{C}$) $\longleftarrow$ VertexTestHat($\mathbf{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T, (\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$)
3: **return** $\boldsymbol{C}(\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$

presentation and development of the algorithm; one can reduce any instance of a partial label model satisfying **(B3)** and **(A)** to an instance of a partial label model that also satisfies **(D)**. We defer the sketch of this reduction to Section E.3.

We now state our main estimation result for classification with partial labels.

Theorem 6 *Let $\delta > 0$ and $\epsilon \in (0, 1)$. Suppose that $P_1, \ldots, P_L$ satisfy **(A″)**, $\mathbf{\Pi}$ has full rank, the columns of $\mathbf{\Pi}^+$ are unique and $\mathbf{\Pi}^+$ satisfies **(D)**. Then, with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, PartialLabelHat($\mathbf{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T \,|\, \epsilon$) returns $(\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$ such that for every $i \in [L]$,*

$$\sup_{E \in \mathcal{E}} |\widehat{Q}_i(E) - P_i(E)| < \delta.$$

5.4. Sieve Estimators

In the preceding, we have assumed a fixed VC class to simplify the presentation. However, these results easily extend to the setting where $\mathcal{E} = \mathcal{E}_k$ and $k \longrightarrow \infty$ at a suitable rate depending on the growth of the VC dimensions V_k. This allows for the P_is to be estimated uniformly on arbitrarily complex events, e.g., $\mathcal{E}_k$ is the set of unions of k open balls.

6. Discussion

In this paper, we have studied the problem of how to recover the base distributions $\boldsymbol{P}$ from the contaminated distributions $\tilde{\boldsymbol{P}}$ without knowledge of the mixing matrix $\mathbf{\Pi}$. We used a common set of concepts and techniques to solve three popular machine learning problems that arise in this setting: multiclass classification with label noise, demixing of mixed membership models, and classification with partial labels. Our technical contributions include: (i) We provide sufficient and sometimes necessary conditions for identifiability for all three problems. (ii) We give nonparametric algorithms for the infinite and finite sample settings. (iii) We provide a new estimator for iterative applications of κ^* that is of independent interest. (iv) Finally, our work provides a novel geometric perspective on each of the three problems.

Our results improve on what was previously known for all three problems. For multiclass classification with label noise and unknown $\mathbf{\Pi}$, previous work had only considered the case $M = L = 2$. Our work achieves a generalization to arbitrarily many distributions. For demixing of mixed membership models, previous algorithms with theoretical guarantees required a finite sample space. Our work allows for a much more general set of distributions. Finally, for classification with partial labels, previous work on learnability assumed the realizable case (non-overlapping $P_1, \ldots, P_L$) and assumed strong conditions on label co-occurence in partial labels. Our analysis covers the agnostic case and a much wider set of partial labels.

Our work has also highlighted the advantages and disadvantages associated with the two-sample κ^* operator and multi-sample κ^* operator, respectively. Algorithms that only use the two-sample κ^* operator have the following two advantages: (i) the geometry of the two-sample κ^* operator is simpler than the geometry of the multi-sample κ^* operator and, as such, can be more tractable. Indeed, in recent years, several practical algorithms for estimating the two-sample κ^* have been developed (see Jain et al. (2016) and references therein). (ii) We have estimators with established rates of convergence for the two-sample κ^* operator, but not for the multi-sample κ^* operator. On the other hand, algorithms that use the multi-sample κ^* operator have fewer steps.

The aims of this work are mainly theoretical, but we believe that our work can inform practical algorithms. First, we note that while we have emphasized recovery of $\boldsymbol{P}$, another interpretation is that our work deals with estimating $\boldsymbol{\Pi}$. One can then plug our estimate of $\boldsymbol{\Pi}$ into corrected losses for multiclass classification with label noise and classification with partial labels that require knowledge of $\boldsymbol{\Pi}$ (Cid-Sueiro, 2012; Menon et al., 2015b; van Rooyen and Williamson, 2015; Patrini et al., 2017). Thus, in general, our work can be applied in this two-stage approach. Second, when L or M are small, the Demix algorithm is practical. For example, Metodiev and Thaler (2018b) apply the Demix algorithm to a high-energy physics application where $L = M = 2$. It is of interest to examine more generally whether variants of Demix work when M or L are small. Third, we conjecture that our analysis suggests novel principles for designing algorithms. For example, an alternative approach to the three problems in question is to embed the contaminated distributions in a reproducing kernel Hilbert Space and to estimate the $\boldsymbol{\Pi}$ matrix by setting up an optimization problem (e.g., see Ramaswamy et al. (2016) for the setting where there are two distributions). One of our necessary conditions, maximality (see Appendix C), suggests formulating the optimization problem to search for base distributions that *(i)* explain the observed contaminated distributions and *(ii)* whose convex hull is as large as possible. In this way, we believe that our general treatment of these three problems that arise in mutual contamination models provides intuitions that could be useful for designing new algorithms.

Although our experiments in Appendix G suggest that joint irreducibility of the base distributions is a reasonable assumption, it is nevertheless worthwhile to consider what questions arise if the base distributions are not exactly jointly irreducible. We see two possible research directions. First, one could perform a stability analysis: when the base distributions are not jointly irreducible, but are nearly jointly irreducible (in some sense that would need to be defined precisely), does the estimate of $\boldsymbol{\Pi}$ remain close to the true $\boldsymbol{\Pi}$? A second research question is to reinterpret the problem of demixing of mixed membership models as a dimensionality reduction problem. That is, given a large set of distributions, one could seek to represent them as convex combinations of a small set of irreducible base distributions. Then, the challenge would be to define an appropriate measure of approximation quality and to determine whether our approach could be useful for designing a consistent algorithm for the best approximation.

Acknowledgements

We thank the anonymous reviewers for their extremely helpful and insightful comments. This work was supported in part by NSF grants 1422157 and 1838179. Gilles Blanchard

acknowledges support from the DFG, under the Research Unit FOR-1735 "Structural Inference in Statistics - Adaptation and Efficiency", and under the Collaborative Research Center SFB-1294 "Data Assimilation".

Appendix A. Outline of Appendices

To begin, we introduce additional notation for the appendices. In Section C, we discuss how strong our sufficient conditions are and present factorization results that suggest that they are reasonable. In Section D, we give our identifiability analysis of demixing mixed membership models and classification with partial labels. In Section E, we prove our results on the ResidueHat estimator, as well as the finite sample algorithms for demixing mixed membership models and classification with partial labels. In Section F, we state some lemmas from related papers that we use in our arguments.

Appendix B. Notation for Appendices

Let $A \subset \mathbb{R}^d$ be a set. Let $\operatorname{aff} A$ denote the affine hull of A, i.e., $\operatorname{aff} A = \{\sum_{i=1}^K \theta_i \boldsymbol{x}_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K \in A, \sum_{i=1}^K \theta_i = 1\}$. A° denotes the relative interior of A, i.e., $A^\circ = \{x \in A | B(x, r) \cap \operatorname{aff} A \subseteq A \text{ for some } r > 0\}$. Then, ∂A denotes the relative boundary of A, i.e., $\partial A = A \backslash A^\circ$. In addition, let $\|\cdot\|$ denote an arbitrary finite-dimensional norm on $\mathbb{R}^L$. For two vectors, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$, define

$$\min(\boldsymbol{x}^T, \boldsymbol{y}^T) = (\min(x_1, y_1), \ldots, \min(x_K, y_K)).$$

$\boldsymbol{x} \geqslant \boldsymbol{y}$ means $x_i \geqslant y_i \ \forall i \in [K]$.

For distributions $Q_1, \ldots, Q_k$, we use $\operatorname{conv}(Q_1, \ldots, Q_K)^\circ$ to denote the relative interior of their convex hull and have that

$$\operatorname{conv}(Q_1, \ldots, Q_K)^\circ = \{\sum_{i=1}^K \alpha_i Q_i : \alpha_i > 0, \sum_{i=1}^K \alpha_i = 1\}.$$

Note that when $Q_1, \ldots, Q_K$ are discrete distributions, this definition coincides with the definition of the relative interior of a set of Euclidean vectors.

We use the following affine mapping throughout the paper: $m_\nu(\boldsymbol{x}, \boldsymbol{y}) = (1 - \nu)\boldsymbol{x} + \nu \boldsymbol{y}$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^L$ and $\nu \in [0, 1]$. Overloading notation, when Q_1 and Q_2 are distributions, we define $m_\nu(Q_1, Q_2) = (1 - \nu)Q_1 + \nu Q_2$. Note that if $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Delta_L$ and $Q_1 = \boldsymbol{\eta}_1^T \boldsymbol{P}$ and $Q_2 = \boldsymbol{\eta}_2^T \boldsymbol{P}$, then $m_\nu(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ is the mixture proportion for the distribution $m_\nu(Q_1, Q_2)$.

Appendix C. Factorization Results

In this section, we discuss whether our sufficient conditions are necessary. For the problems of demixing mixed membership models and classification with partial labels, we provide factorization results that suggest that our sufficient conditions are not much stronger than what is necessary.

C.1. Multiclass Classification with Label Noise

Our sufficient condition **(B1)** for multiclass classification with label noise is not necessary. Rather, **(B1)** is one of several possible sufficient conditions, and one that reflects a low noise assumption as illustrated in Figure 1. Consider the case $L = M = 2$ where **(B1)** is equivalent to $\pi_{1,2} + \pi_{2,1} < 1$. Recovery is still possible if $\pi_{1,2} + \pi_{2,1} > 1$ since one can simply swap $\tilde{P}_1$ and $\tilde{P}_2$ in a decontamination procedure. $\pi_{1,2} + \pi_{2,1} < 1$ is only necessary if one assumes that most of the training labels are correct, which is what $\pi_{1,2} + \pi_{2,1} < 1$ essentially says. For larger $L = M$, **(B1)** says in a sense that most of the data from $\tilde{P}_i$ come from P_i for every i. Other sufficient conditions are possible (as in the binary case), but these would require at least one $\tilde{P}_i$ to contain a significant portion of some P_j, $j \neq i$. Regarding **(A)**, Blanchard et al. (2016) study the question of necessity for joint irreducibility in the case $L = M = 2$ and show that under mild assumptions on the decontamination procedure, joint irreducibility is necesssary.

C.2. Demixing Mixed Membership Models

Recall the definition of a factorization: $\tilde{\boldsymbol{P}}$ is factorizable if there exists $(\boldsymbol{\Pi}, \boldsymbol{P}) \in \Delta_L^M \times \mathcal{P}^L$ such that $\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}\boldsymbol{P}$; we call $(\boldsymbol{\Pi}, \boldsymbol{P})$ a factorization of $\tilde{\boldsymbol{P}}$.

Our sufficient conditions are not much stronger than what is required by factorizations that satisfy the two forthcoming desirable properties.

Definition 7 *We say a factorization $(\boldsymbol{\Pi}, \boldsymbol{P})$ of $\tilde{\boldsymbol{P}}$ is* maximal ***(M)*** *iff for all factorizations $(\boldsymbol{\Pi}', \boldsymbol{P}')$ of $\tilde{\boldsymbol{P}}$ with $\boldsymbol{P}' = (P_1', \ldots, P_L')^T \in \mathcal{P}^L$, it holds that $\{P_1', \ldots, P_L'\} \subseteq \mathrm{conv}(P_1, \ldots, P_L)$.*

In words, $\boldsymbol{P} = (P_1, \ldots, P_L)^T$ is a maximal collection of base distributions if it is not possible to move any of the P_is outside of $\mathrm{conv}(P_1, \ldots, P_L)$ and represent $\tilde{\boldsymbol{P}}$.

Definition 8 *We say a factorization $(\boldsymbol{\Pi}, \boldsymbol{P})$ of $\tilde{\boldsymbol{P}}$ is* linear ***(L)*** *iff $\{P_1, \ldots, P_L\} \subseteq \mathrm{span}(\tilde{P}_1, \ldots, \tilde{P}_M)$.*

We believe that **(L)** is a reasonable requirement because it holds in the common situation in which there exist $\boldsymbol{\pi}_{i_1}, \ldots, \boldsymbol{\pi}_{i_L}$ that are linearly independent. Then for $I = \{i_1, \ldots, i_L\}$, we can write $\tilde{\boldsymbol{P}}_I = \boldsymbol{\Pi}_I \boldsymbol{P}$ where $\boldsymbol{\Pi}_I$ is the submatrix of $\boldsymbol{\Pi}$ containing only the rows indexed by I and $\tilde{\boldsymbol{P}}_I$ is similarly defined. Then, $\boldsymbol{\Pi}_I$ is invertible and $\boldsymbol{P} = \boldsymbol{\Pi}_I^{-1} \tilde{\boldsymbol{P}}_I$.

Factorizations that satisfy **(A)** and **(B2)** are maximal and linear.

Proposition 5 *Let $(\boldsymbol{\Pi}, \boldsymbol{P})$ be a factorization of $\tilde{\boldsymbol{P}}$. If $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(A)** and **(B2)**, then $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(M)** and **(L)**.*

Proof We first show that $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(L)**. By hypothesis, $P_1, \ldots, P_L$ are jointly irreducible. By Lemma 16, $P_1, \ldots, P_L$ are linearly independent. Since by hypothesis $\boldsymbol{\Pi}$ has full rank, there exist L rows in $\boldsymbol{\Pi}$, $\boldsymbol{\pi}_{i_1}, \ldots, \boldsymbol{\pi}_{i_L}$, that are linearly independent. By Lemma 16, $\tilde{P}_{i_1}, \ldots, \tilde{P}_{i_L}$ are linearly independent. Since $\boldsymbol{\Pi}\boldsymbol{P} = \tilde{\boldsymbol{P}}$, $\mathrm{span}(\tilde{P}_1, \ldots, \tilde{P}_M) \subseteq \mathrm{span}(P_1, \ldots, P_L)$. Since $\dim \mathrm{span}(\tilde{P}_1, \ldots, \tilde{P}_M) \geqslant L$, we have $\mathrm{span}(\tilde{P}_1, \ldots, \tilde{P}_M) = \mathrm{span}(P_1, \ldots, P_L)$. Therefore, $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(L)**.

Now, we show that $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(M)**. Suppose that there is another solution $(\boldsymbol{\Pi}', \boldsymbol{P}')$ with $\boldsymbol{P}' = (P_1', \ldots, P_L')^T$ such that $\boldsymbol{\Pi}'\boldsymbol{P}' = \tilde{\boldsymbol{P}}$ and with some P_i' such that

$P_i' \notin \text{conv}(P_1, \ldots, P_L)$. We claim that $P_i' \notin \text{span}(P_1, \ldots, P_L)$. Towards a contradiction, suppose that $P_i' \in \text{span}(P_1, \ldots, P_L)$ so that we can write $P_i' = \sum_{i=1}^{L} a_i P_i$. Then, at least one of the a_i is negative since, by assumption, $P_i' \notin \text{conv}(P_1, \ldots, P_L)$. But, by joint irreducibility of $P_1, \ldots, P_L$, P_i' is not a distribution, which is a contradiction. So, the claim follows. But, then, since $\text{span}(P_1, \ldots, P_L) = \text{span}(\tilde{P}_1, \ldots, \tilde{P}_M)$, we must have that $\text{span}(\tilde{P}_1, \ldots, \tilde{P}_M) \subseteq \text{span}(P_1', \ldots, P_{i-1}', P_{i+1}', \ldots, P_L)$, which is impossible since $\dim \text{span}(\tilde{P}_1, \ldots, \tilde{P}_M) = L$. ■

Maximal and linear factorizations imply conditions that are not much weaker than our sufficient conditions.

Theorem 7 *Let* $(\boldsymbol{\Pi}, \boldsymbol{P})$ *be a factorization of* $\tilde{\boldsymbol{P}}$. *If* $(\boldsymbol{\Pi}, \boldsymbol{P})$ *satisfies* ***(M)****, then*

(A′) $\forall i$, P_i *is irreducible with respect to every distribution in* $\text{conv}(\{P_j : j \neq i\})$.

If $(\boldsymbol{\Pi}, \boldsymbol{P})$ *satisfies* ***(L)****, then*

(B′) $\text{rank}(\boldsymbol{\Pi}) \geqslant \dim \text{span}(P_1, \ldots, P_L)$.

Proof

(A′) We prove the contrapositive. Suppose that there is some P_i and $Q \in \text{conv}(\{P_j : j \neq i\})$ with $Q = \sum_{j \neq i} \beta_j P_j$ such that P_i is not irreducible wrt Q. Then, there is some distribution G and $\gamma \in (0, 1]$ such that $P_i = \gamma Q + (1 - \gamma) G$.

Suppose $\gamma = 1$. Then, $P_i = Q \in \text{conv}(\{P_k : k \neq i\})$. But, then $\tilde{P}_1, \ldots, \tilde{P}_M \in \text{conv}(\{P_j : j \neq i\} \cup \{R\})$ for any distribution $R \notin \text{conv}(P_1, \ldots, P_L)$. This shows that $(\boldsymbol{\Pi}, \boldsymbol{P})$ does not satisfy **(M)**.

Therefore, assume that $\gamma \in (0, 1)$. Either $G \in \text{conv}(P_1, \ldots, P_L)$ or $G \notin \text{conv}(P_1, \ldots, P_L)$. Suppose that $G \in \text{conv}(P_1, \ldots, P_L)$. Then, there exist $\alpha_1, \ldots, \alpha_L$ all nonnegative and summing to 1 such that

$$P_i = \gamma Q + (1 - \gamma)(\alpha_1 P_1 + \ldots + \alpha_L P_L).$$

Therefore, $P_i \in \text{conv}(\{P_k : k \neq i\})$. Then, by the argument in the previous paragraph, $(\boldsymbol{\Pi}, \boldsymbol{P})$ does not satisfy **(M)**.

Now, suppose that $G \notin \text{conv}(P_1, \ldots, P_L)$. Since $P_i \in \text{conv}(G, Q)$ and $Q \in \text{conv}(\{P_j : j \neq i\})$, we have that $\text{conv}(\{P_j : j \neq i\} \cup \{G\}) \supset \text{conv}(P_1, \ldots, P_L)$. Then, $\tilde{P}_1, \ldots, \tilde{P}_M \in \text{conv}(\{P_j : j \neq i\} \cup \{G\})$. This shows that $(\boldsymbol{\Pi}, \boldsymbol{P})$ does not satisfy **(M)**. The result follows.

(B′) Clearly, $\dim \text{span}(\tilde{P}_1, \ldots, \tilde{P}_M) \leqslant \dim \text{span}(P_1, \ldots, P_L)$ since $\tilde{P}_i = \boldsymbol{\pi}_i^T \boldsymbol{P}$ for all $i \in [M]$. Since $(\boldsymbol{\Pi}, \boldsymbol{P})$ satisfies **(L)**, $\text{span}(P_1, \ldots, P_L) \subset \text{span}(\tilde{P}_1, \ldots, \tilde{P}_L)$, which implies that

$$\dim \text{span}(P_1, \ldots, P_L) \leqslant \dim \text{span}(\tilde{P}_1, \ldots, \tilde{P}_M).$$

Therefore, $\dim \text{span}(\tilde{P}_1, \ldots, \tilde{P}_M) = \dim \text{span}(P_1, \ldots, P_L)$. Then, since $\tilde{P}_1, \ldots, \tilde{P}_M \in \text{range}(\boldsymbol{\Pi})$, $\dim \text{span}(P_1, \ldots, P_L) \leqslant \dim \text{range}\, \boldsymbol{\Pi}$. By Result 3.117 of Axler (2015),

$$\text{rank}(\boldsymbol{\Pi}) = \dim \text{range}\, \boldsymbol{\Pi} \geqslant \dim \text{span}(P_1, \ldots, P_L).$$

■

As a corollary, Theorem 7 implies that if there is a linear factorization $(\mathbf{\Pi}, \boldsymbol{P})$ of $\tilde{P}$ and $P_1, \ldots, P_L$ are linearly independent, then there must be at least as many contaminated distributions as base distributions, i.e., $M \geqslant L$. Also, note that **(A′)** appears as a sufficient condition in Sanderson and Scott (2014).

By comparing **(A)** with **(A′)** and **(B2)** with **(B′)**, we see that the proposed sufficient conditions are not much stronger than **(M)** and **(L)** require. Since joint irreducibility of $P_1, \ldots, P_L$ entails their linear independence by Lemma 16, under **(A)**, **(B2)** and **(B′)** are the same. **(A)** differs from **(A′)** in that it requires that a slightly larger set of distributions are irreducible with respect to convex combinations of the remaining distributions. Specifically, under **(A)**, *every convex combination* of a subset of the P_is is irreducible with respect to every convex combination of the other P_is whereas **(A′)** only requires that every P_i be irreducible with respect to every convex combination of the other P_is.

C.3. Classification with Partial Labels

Most of our definitions and results for classification with partial labels parallel those of demixing mixed membership models. We say that $\tilde{\boldsymbol{P}}$ is $\mathbf{\Pi}^+$-factorizable if there exists a pair $(\mathbf{\Pi}, \boldsymbol{P}) \in \Delta_L^M \times \mathcal{P}^L$ that solves (2) such that $\mathbf{\Pi}$ is consistent with $\mathbf{\Pi}^+$; we call $(\mathbf{\Pi}, \boldsymbol{P})$ an $\mathbf{\Pi}^+$-factorization of $\tilde{\boldsymbol{P}}$. We say a partial label model is identifiable if given $(\tilde{\boldsymbol{P}}, \mathbf{\Pi}^+)$, $\tilde{\boldsymbol{P}}$ has a unique $\mathbf{\Pi}^+$-factorization $(\mathbf{\Pi}, \boldsymbol{P})$.

Our definitions of maximal and linear $\mathbf{\Pi}^+$-factorizations resemble definitions 7 and 8.

Definition 9 *We say a $\mathbf{\Pi}^+$-factorization $(\mathbf{\Pi}, \boldsymbol{P})$ of $\tilde{\boldsymbol{P}}$ is* maximal ***(M)*** *iff for all $\mathbf{\Pi}^+$-factorizations $(\mathbf{\Pi}', \boldsymbol{P}')$ of $\tilde{\boldsymbol{P}}$ with $\boldsymbol{P}' = (P_1', \ldots, P_L')^T \in \mathcal{P}^L$, it holds that $\{P_1', \ldots, P_L'\} \subseteq \operatorname{conv}(P_1, \ldots, P_L)$.*

Definition 10 *We say a $\mathbf{\Pi}^+$-factorization $(\mathbf{\Pi}, \boldsymbol{P})$ of $\tilde{\boldsymbol{P}}$ is* linear ***(L)*** *iff $\{P_1, \ldots, P_L\} \subseteq \operatorname{span}(\tilde{P}_1, \ldots, \tilde{P}_M)$.*

Similarly, $\mathbf{\Pi}^+$-factorizations that satisfy **(A)** and **(B3)** are maximal and linear. The proof is identical and, accordingly, omitted.

Proposition 6 *Let $(\mathbf{\Pi}, \boldsymbol{P})$ be a $\mathbf{\Pi}^+$-factorization of $\tilde{\boldsymbol{P}}$. If $(\mathbf{\Pi}, \boldsymbol{P})$ satisfies **(A)** and **(B3)**, then $(\mathbf{\Pi}, \boldsymbol{P})$ satisfies **(M)** and **(L)**.*

Linear $\mathbf{\Pi}^+$-factorizations must satisfy **(B′)**; indeed, the proof is identical to the proof for linear factorizations. However, maximal $\mathbf{\Pi}^+$-factorizations need not satisfy **(A′)**. Consider the following counterexample. Let $Q_1 \sim \text{unif}(0, 2)$ and $Q_2 \sim \text{unif}(1, 3)$. Let $P_1 = \frac{2}{3}Q_1 + \frac{1}{3}Q_2$, $P_2 = \frac{1}{3}Q_1 + \frac{2}{3}Q_2$, $\tilde{P}_1 = P_1$, and $\tilde{P}_2 = P_2$. Then, $\mathbf{\Pi}^+ = \boldsymbol{I}_2$—the identity matrix. Then, any $(\mathbf{\Pi}', \boldsymbol{P}')$ that satisfies (2) and is consistent with $\mathbf{\Pi}^+$ must be such that $(P_1, P_2)^T = \boldsymbol{P}'$. Therefore, **(M′)** is satisfied. But, clearly, **(A′)** is not satisfied.

In summary, we are unable to offer a necessary condition that is close to **(A)**. On the other hand, **(B3)** is necessary.

Proposition 7 *Let $(\mathbf{\Pi}, \boldsymbol{P})$ be an $\mathbf{\Pi}^+$-factorization of $\tilde{\boldsymbol{P}}$. If $(\tilde{\boldsymbol{P}}, \mathbf{\Pi}^+)$ is identifiable, then the columns of $\mathbf{\Pi}^+$ are distinct.*

Proof First, suppose $(\tilde{\boldsymbol{P}}, \boldsymbol{\Pi}^+)$ is identifiable. Then, we can write $\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}\boldsymbol{P}$ where $\boldsymbol{\Pi}$ is consistent with $\boldsymbol{\Pi}^+$. We claim that for all $i \neq j$, $P_i \neq P_j$. To the contrary, suppose that there exists $i \neq j$ such that $P_i = P_j$. Without loss of generality, suppose $i = 1, j = 2$. Then, we can write

$$\tilde{\boldsymbol{P}} = \begin{pmatrix} 2\boldsymbol{\Pi}_{:,1} & \boldsymbol{\Pi}_{:,3} & \dots & \boldsymbol{\Pi}_{:,L} \end{pmatrix} \begin{pmatrix} P_1 \\ P_3 \\ \vdots \\ P_L \end{pmatrix},$$

which contradicts the uniqueness of $\boldsymbol{P}$ and $\boldsymbol{\Pi}$.

Next, we give a proof by contraposition. Suppose that there exists $i \neq j$ such that $\boldsymbol{\Pi}^+_{:,i} = \boldsymbol{\Pi}^+_{:,j}$. Without loss of generality, let $i = 1$ and $j = 2$. Suppose that $(\boldsymbol{\Pi}, \boldsymbol{P})$ is consistent with $\boldsymbol{\Pi}^+$ and solves $\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}\boldsymbol{P}$. Then, the pair $(\boldsymbol{\Pi}', \boldsymbol{P}')$ given by

$$\boldsymbol{\Pi}' = \begin{pmatrix} \boldsymbol{\Pi}_{:,2} & \boldsymbol{\Pi}_{:,1} & \boldsymbol{\Pi}_{:,3} & \dots & \boldsymbol{\Pi}_{:,L} \end{pmatrix}$$

$$\boldsymbol{P}' = \begin{pmatrix} P_2 \\ P_1 \\ P_3 \\ \vdots \\ P_L \end{pmatrix}$$

solves $\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}'\boldsymbol{P}'$ and is consistent with $\boldsymbol{\Pi}^+$. If $P_1 = P_2$, then $(\tilde{\boldsymbol{P}}, \boldsymbol{\Pi}^+)$ is not identifiable, so we may rule out this case. Therefore, $\boldsymbol{P}' \neq \boldsymbol{P}$, yielding the result. ■

Appendix D. Identification

In this section, we establish our identification results, i.e., Theorems 2 and 3. We begin by proving Proposition 3. Second, we prove a set of useful lemmas in Section D.2. Third, we present our results on demixing mixed membership models in Section D.3. Finally, we present our results on classification with partial labels in Section D.4.

D.1. Proof of Proposition 3

Proof We prove the claims in order.

1. Without loss of generality, suppose $i = 1$ and let $A = [L]\backslash\{1\}$. There is at least one point attaining the maximum in the optimization problem $\kappa^*(\boldsymbol{\eta}_1 \,|\, \{\boldsymbol{\eta}_j : j \neq 1\})$ by Lemma 15. Take a G that achieves the maximum in $\kappa^*(Q_1 \,|\, \{Q_j : j \neq 1\})$, which exists also by Lemma 15. Then, we can write:

$$Q_1 = (1 - \sum_{j \geqslant 2} \mu_j)G + \sum_{j \geqslant 2} \mu_j Q_j. \tag{7}$$

   Note that since $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_L$ are linearly independent and $P_1, \dots, P_L$ are jointly irreducible, $Q_1, \dots, Q_L$ are linearly independent by Lemma 16. Therefore, $\kappa^*(Q_1 \,|\, \{Q_j : j \neq 1\}) = \sum_{j \geqslant 2} \mu_j < 1$ because, if not, $Q_1 = \sum_{j \geqslant 2} \mu_j Q_j$.

Further, any G that satisfies (7) has the form $\sum_{i=1}^{L} \gamma_i P_i$ because (7) implies that $G \in \text{span}(Q_1, \ldots, Q_L)$ and each $Q_i \in \text{conv}(P_1, \ldots, P_L)$ by hypothesis. The γ_i must sum to one, and we have that they are nonnegative by joint irreducibility. That is, $\boldsymbol{\gamma} \equiv \left(\gamma_1, \ldots, \gamma_L\right)^T$ is a discrete distribution. Then, the above equation is equivalent to

$$\boldsymbol{\eta}_1^T \boldsymbol{P} = (1 - \sum_{j \geqslant 2} \mu_j) \boldsymbol{\gamma}^T \boldsymbol{P} + \sum_{j \geqslant 2} \mu_j \boldsymbol{\eta}_j^T \boldsymbol{P}. \tag{8}$$

Since $P_1, \ldots, P_L$ are jointly irreducible, $P_1, \ldots, P_L$ are linearly independent by Lemma 16. By linear independence of $P_1, \ldots, P_L$, we obtain

$$\boldsymbol{\eta}_1 = (1 - \sum_{j \geqslant 2} \mu_j) \boldsymbol{\gamma} + \sum_{j \geqslant 2} \mu_j \boldsymbol{\eta}_j. \tag{9}$$

Consequently, $\kappa^*(Q_1 \,|\, \{Q_j : j \neq 1\}) = \kappa^*(\boldsymbol{\eta}_1 \,|\, \{\boldsymbol{\eta}_j : j \neq 1\}) < 1$. This completes the proof of statement 1.

2. This result follows immediately from Lemma 15.

3. By equations (8) and (9), there is a one-to-one correspondence between the maximizer G to $\kappa^*(Q_1 \,|\, \{Q_j : j \neq 1\})$ and the maximizer $\boldsymbol{\gamma}$ to $\kappa^*(\boldsymbol{\eta}_1 \,|\, \{\boldsymbol{\eta}_j : j \neq 1\})$. The one-to-one correspondence is given by $G = \boldsymbol{\gamma}^T \boldsymbol{P}$.

■

D.2. Lemmas for Identification

We present some technical results that are used repeatedly for our identification results. Lemma 2 gives us some useful properties of the two-sample κ^* that we exploit in the PartialLabel and Demix algorithms. Statement *1* gives an alternative form of κ^*. Statement *2* gives the intuitive result that the residues lie on the boundary of the simplex. Statement *3* gives a useful relation for determining whether two mixture proportions are on the same face; we use this relation extensively in our algorithms.

Lemma 2 *Let $F_1, \ldots, F_K$ be jointly irreducible distributions with $\boldsymbol{F} = (F_1, \ldots, F_K)^T$, Q_1, Q_2 be two distributions such that $Q_i = \boldsymbol{\eta}_i^T \boldsymbol{F}$ where $\boldsymbol{\eta}_i \in \Delta_K$ for $i = 1, 2$ and $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$. Let R be the residue of Q_1 wrt Q_2 and $R = \boldsymbol{\mu}^T \boldsymbol{F}$.*

1. *There is a one-to-one correspondence between the optimization problem in $\kappa^*(Q_1 \,|\, Q_2)$ and the optimization problem*

$$\max(\alpha \geqslant 1 | \exists \text{ a distribution } G \text{ s.t } G = Q_2 + \alpha(Q_1 - Q_2))$$

via $\alpha = (1 - \kappa)^{-1}$.

2. *$\boldsymbol{\mu} \in \partial \Delta_K$.*

3. *$\mathcal{N}(\boldsymbol{\eta}_2) \nsubseteq \mathcal{N}(\boldsymbol{\eta}_1)$ if and only if $R = Q_1$ if and only if $\kappa^*(Q_1 \,|\, Q_2) = 0$.*

Proof We note that we may assume that $R = \boldsymbol{\mu}^T \boldsymbol{F}$ since by definition of the residue, $R \in \text{span}(Q_1, Q_2)$ and $Q_1, Q_2 \in \text{conv}(F_1, \ldots, F_K)$.

1. Consider the linear relation: $Q_1 = (1-\kappa)G + \kappa Q_2$ where $\kappa \in [0,1]$. Since $F_1, \ldots, F_K$ are jointly irreducible and $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are linearly independent, Q_1 and Q_2 are linearly independent by Lemma 16. Therefore, $\kappa < 1$. We can rewrite the relation as

$$G = \frac{1}{1-\kappa} Q_1 - \frac{\kappa}{1-\kappa} Q_2 = \alpha Q_1 + (1-\alpha) Q_2$$

   where $\alpha = \frac{1}{1-\kappa}$. The equivalence follows.

2. Since R is the residue of Q_1 wrt Q_2, by Proposition 3, $\boldsymbol{\mu}$ is the residue of $\boldsymbol{\eta}_1$ wrt $\boldsymbol{\eta}_2$ and $\boldsymbol{\mu} \in \Delta_K$. Therefore, by statement *1* in Lemma 2, $\boldsymbol{\mu}$ is such that α^* is maximized subject to the following constraints:

$$\begin{aligned} \boldsymbol{\mu} &= (1-\alpha^*)\boldsymbol{\eta}_2 + \alpha^* \boldsymbol{\eta}_1 \\ \alpha^* &\geqslant 1 \\ \boldsymbol{\mu} &\in \Delta_K. \end{aligned}$$

   Suppose that $\min_i \mu_i > 0$. Then,

$$\boldsymbol{\mu} = (1-\alpha^*)\boldsymbol{\eta}_2 + \alpha^* \boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 + \alpha(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) > 0$$

   so that there is some $\epsilon > 0$ such that

$$\begin{aligned} \boldsymbol{\mu}' &= (1-\alpha^*-\epsilon)\boldsymbol{\eta}_2 + (\alpha^*+\epsilon)\boldsymbol{\eta}_1 \\ \alpha^* + \epsilon &\geqslant 1 \\ \boldsymbol{\mu}' &\in \Delta_K. \end{aligned}$$

   But, this contradicts the definition of α^* and $\boldsymbol{\mu}$. Therefore, $\min_i \mu_i = 0$. Consequently, $\boldsymbol{\mu} \in \partial\Delta_K$.

3. By definition of κ^*, it is clear that $R = Q_1$ if and only if $\kappa^*(Q_1 \,|\, Q_2) = 0$. Therefore, it suffices to show that $\mathcal{N}(\boldsymbol{\eta}_2) \nsubseteq \mathcal{N}(\boldsymbol{\eta}_1)$ if and only if $\kappa^*(Q_1 \,|\, Q_2) = 0$. Suppose $\mathcal{N}(\boldsymbol{\eta}_2) \nsubseteq \mathcal{N}(\boldsymbol{\eta}_1)$. Then, there must be $i \in [K]$ such that $\eta_{2,i} > 0$ and $\eta_{1,i} = 0$. For any $\alpha > 1$,

$$\min_{i \in [K]} (1-\alpha)\eta_{2,i} + \alpha \eta_{1,i} < 0;$$

   but, this violates the constraint of the optimization problem. Therefore, $\alpha = 1$. By statement *1* in Lemma 2, $\kappa^*(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) = 0$. By statement *1* of Proposition 3, $\kappa^*(Q_1 \,|\, Q_2) = \kappa^*(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) = 0$.

   Now, suppose $\mathcal{N}(\boldsymbol{\eta}_2) \subseteq \mathcal{N}(\boldsymbol{\eta}_1)$. Then, for any $i \in [K]$, if $\eta_{2,i} > 0$, then $\eta_{1,i} > 0$. Then, there is $\alpha > 1$ sufficiently close to 1 such that

$$\min_{i \in [K]} \eta_{2,i} + \alpha(\eta_{1,i} - \eta_{2,i}) \geqslant 0.$$

   By statement *1* in this Lemma, $\kappa^*(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) > 0$. By statement *1* of proposition 3, $\kappa^*(Q_1 \,|\, Q_2) = \kappa^*(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) > 0$.

■

Lemma 3 *Let $0 \leqslant k < L$. If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \Delta_L$ are linearly independent and $\boldsymbol{w}_{k+1}, \ldots, \boldsymbol{w}_L \in \Delta_L$ are random vectors drawn independently from the uniform distribution on a set $A \subset \Delta_L$ with positive $(L-1)$-dimensional Lebesgue measure, then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \boldsymbol{w}_{k+1}, \ldots, \boldsymbol{w}_L$ are linearly independent with probability* 1.

Proof We prove the result inductively. To begin, we prove the base case, i.e. $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \boldsymbol{w}_{k+1}$ are linearly independent w.p. 1. It suffices to show that $\boldsymbol{w}_{k+1} \notin \text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ w.p. 1. Thus, it is enough to show that $\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cap A$ has $(L-1)$-dimensional Lebesgue measure 0. Since

$$\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cap A \subset \text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cap \Delta_L,$$

it suffices to show that $\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cap \Delta_L$ has $(L-1)$-dimensional Lebesgue measure 0.

Next, we claim that $\text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cap \Delta_L \subseteq \text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. Let $\sum_{i=1}^k \alpha_i \boldsymbol{v}_i \in \Delta_L$. Since $\boldsymbol{v}_i \in \Delta_L$ for all $i \in [k]$, we can write $\boldsymbol{v}_i = \sum_{j=1}^L \beta_{i,j} \boldsymbol{e}_i$ where $\beta_{i,j} \geqslant 0$ and $\sum_{j=1}^L \beta_{i,j} = 1$. Then, since $\sum_{i=1}^k \alpha_i \boldsymbol{v}_i = \sum_{i=1}^k \alpha_i \sum_{j=1}^L \beta_{i,j} \boldsymbol{e}_j \in \Delta_L$, it holds that

$$1 = \sum_{i=1}^k \alpha_i \sum_{j=1}^L \beta_{i,j} = \sum_{i=1}^k \alpha_i.$$

Thus, $\sum_{i=1}^k \alpha_i \boldsymbol{v}_i \in \text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$, establishing the claim.

Thus, it suffices to show that $\text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ has $(L-1)$-dimensional Lebesgue measure 0. $\text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ has affine dimension at most $k-1$. Since it is not possible to fit a $(L-1)$ dimensional ball in an affine subspace of affine dimension $k-1 < L-1$, $\text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ has $(L-1)$-dimensional Lebesgue measure 0. Thus, with probability 1, $\boldsymbol{w}_{k+1} \notin \text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. This establishes the base case.

The inductive step follows by a union bound and a similar argument to the base case. Thus, the result follows. ■

D.3. Demixing Mixed Membership Models

In this section, we prove our identification result for demixing mixed membership models, i.e., Theorem 2. First, we present technical lemmas in Section D.3.1. Second, in Section D.3.2, we present the key subroutine FaceTest and prove that it behaves as desired. Third, we prove Theorem 2 in Section D.3.3. Finally, in Section D.3.4, we extend our results to the nonsquare case (where $M > L$).

D.3.1. Lemmas

Lemma 4 establishes an intuitive continuity property of the two-sample version of κ^* and the residue. Recall that $\|\cdot\|$ denotes an arbitrary finite-dimensional norm on $\mathbb{R}^L$.

Lemma 4 *Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Delta_L$ be distinct vectors and let $\boldsymbol{\mu}$ be the residue of $\boldsymbol{\eta}_2$ wrt $\boldsymbol{\eta}_1$. Let $\boldsymbol{\gamma}_n \in \Delta_L$ be a sequence such that $\|\boldsymbol{\gamma}_n - \boldsymbol{\eta}_2\| \longrightarrow 0$ as $n \longrightarrow \infty$, and let $\boldsymbol{\tau}_n$ be the residue of $\boldsymbol{\gamma}_n$ wrt $\boldsymbol{\eta}_1$. Then,*

1. $\lim_{n\longrightarrow\infty} \kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\eta}_1) = \kappa^*(\boldsymbol{\eta}_2 \,|\, \boldsymbol{\eta}_1)$, *and*
2. $\lim_{n\longrightarrow\infty} \|\boldsymbol{\tau}_n - \boldsymbol{\mu}\| = 0$.
3. *If, in addition, $\boldsymbol{\rho}_n \in \Delta_L$ is a sequence such that $\|\boldsymbol{\rho}_n - \boldsymbol{\eta}_2\| \longrightarrow 0$ as $n \longrightarrow \infty$ and $\mathcal{N}(\boldsymbol{\eta}_2) = \mathcal{N}(\boldsymbol{\gamma}_n) = \mathcal{N}(\boldsymbol{\rho}_n)$ for all n. Then, $\lim_{n\longrightarrow\infty} \kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\rho}_n) = 1$.*

Proof

1. In order to apply the residue operator κ^* to $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\gamma}_n$ we think of $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\gamma}_n$ as discrete probability distributions. By Proposition 2,

$$\kappa^*(\boldsymbol{\eta}_2 \,|\, \boldsymbol{\eta}_1) = \min_{i,\, \eta_{1,i}>0} \frac{\eta_{2,i}}{\eta_{1,i}}.$$

   Clearly, there is a constant $\delta > 0$ such that $\min_{i,\, \boldsymbol{\eta}_{1,i}>0} \boldsymbol{\eta}_{1,i} > \delta$. Let $\epsilon > 0$. By the equivalence of norms on finite-dimensional vector spaces, there exists a constant $C > 0$ such that $\|\cdot\|_\infty \leqslant C\,\|\cdot\|$ where $\|\cdot\|_\infty$ denotes the supremum norm. Thus, since $\|\boldsymbol{\gamma}_n - \boldsymbol{\eta}_2\| \longrightarrow 0$ as $n \longrightarrow \infty$, we can let n large enough such that $|\boldsymbol{\gamma}_{n,i} - \boldsymbol{\eta}_{2,i}| \leqslant \epsilon$ for all $i \in [L]$. Then,

$$\begin{aligned}\kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\eta}_1) = \min_{i,\eta_{1,i}>0} \frac{\gamma_{n,i}}{\eta_{1,i}} &\leqslant \min_{i,\eta_{1,i}>0} \frac{\eta_{2,i}+\epsilon}{\eta_{1,i}} \\ &\leqslant \kappa^*(\boldsymbol{\eta}_2 \,|\, \boldsymbol{\eta}_1) + \frac{\epsilon}{\delta}.\end{aligned}$$

   Similarly,

$$\begin{aligned}\kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\eta}_1) &= \min_{i,\eta_{1,i}>0} \frac{\gamma_{n,i}}{\eta_{1,i}} \\ &\geqslant \frac{\eta_{2,i}-\epsilon}{\eta_{1,i}} \\ &\geqslant \kappa^*(\boldsymbol{\eta}_2 \,|\, \boldsymbol{\eta}_1) - \frac{\epsilon}{\delta}.\end{aligned}$$

   Since $\epsilon > 0$ was arbitrary, statement *1* follows.

2. Write $\boldsymbol{\mu} = \kappa\boldsymbol{\eta}_2 + (1-\kappa)\boldsymbol{\eta}_1$ and $\boldsymbol{\tau}_n = \kappa_n\boldsymbol{\gamma}_n + (1-\kappa_n)\boldsymbol{\eta}_1$ where $\kappa = \kappa^*(\boldsymbol{\eta}_2 \,|\, \boldsymbol{\eta}_1)$ and $\kappa_n = \kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\eta}_1)$. Then, by the triangle inequality,

$$\begin{aligned}\|\boldsymbol{\mu} - \boldsymbol{\tau}_n\| &\leqslant \|\kappa\boldsymbol{\eta}_2 - \kappa_n\boldsymbol{\gamma}_n\| + \|(1-\kappa)\boldsymbol{\eta}_1 - (1-\kappa_n)\boldsymbol{\eta}_1\| \\ &\leqslant |\kappa-\kappa_n|\,\|\boldsymbol{\eta}_2\| + |\kappa_n|\,\|\boldsymbol{\eta}_2 - \boldsymbol{\gamma}_n\| + |\kappa-\kappa_n|\,\|\boldsymbol{\eta}_1\| \longrightarrow 0\end{aligned}$$

   as $n \longrightarrow \infty$.

3. W.l.o.g., suppose that $[K] = \mathcal{N}(\boldsymbol{\eta}_2) = \mathcal{N}(\boldsymbol{\gamma}_n) = \mathcal{N}(\boldsymbol{\rho}_n)$ where $K \leqslant L$. Observe that

$$\kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\rho}_n) = \min_{i,\, \rho_{n,i}>0} \frac{\gamma_{n,i}}{\rho_{n,i}} = \min_{i \in [K]} \frac{\gamma_{n,i}}{\rho_{n,i}}$$

There exists a constant $\delta > 0$ such that $\min_{i\in[K]} \eta_{2,i} \geqslant \delta$. Let $\delta > \epsilon > 0$. By the equivalence of norms on finite-dimensional vector spaces, we can let n large enough such that $|\gamma_{n,i} - \eta_{2,i}| \leqslant \epsilon$ and $|\rho_{n,i} - \eta_{2,i}| \leqslant \epsilon$ for all $i \in [L]$. Then,

$$\begin{aligned}
\kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\rho}_n) &= \min_{i \in [K]} \frac{\gamma_{n,i}}{\rho_{n,i}} \\
&\leqslant \min_{i \in [K]} \frac{\eta_{2,i} + \epsilon}{\eta_{2,i} - \epsilon} \\
&\leqslant \frac{\eta_{2,i} + \epsilon}{\eta_{2,i} - \epsilon}
\end{aligned}$$

for any $i \in [K]$, which goes to 1 as $\epsilon \longrightarrow 0$. Similarly,

$$\begin{aligned}
\kappa^*(\boldsymbol{\gamma}_n \,|\, \boldsymbol{\rho}_n) &= \min_{i \in [K]} \frac{\gamma_{n,i}}{\rho_{n,i}} \\
&\geqslant \min_{i \in [K]} \frac{\eta_{2,i} - \epsilon}{\eta_{2,i} + \epsilon}
\end{aligned}$$

Since for any $i \in [K]$,

$$\frac{\eta_{2,i} - \epsilon}{\eta_{2,i} + \epsilon} \longrightarrow 1$$

as $\epsilon \longrightarrow 0$, the above lower bound goes to 1 as $\epsilon \longrightarrow 0$. Thus, statement *3* follows.

■

Lemma 5 guarantees that certain operations in the Demix algorithm preserve linear independence of the mixture proportions. The proof uses tools from multilinear algebra.

Lemma 5 *Let $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K \in \Delta_K$ be linearly independent and $P_1, \ldots, P_K$ be jointly irreducible. Let $Q_i = \boldsymbol{\tau}_i^T \boldsymbol{P}$ for $i \in [K]$. Then for any $i, j \in [K]$ such that $i \neq j$,*

1. *If $\boldsymbol{\eta} = \sum_{k=1}^K a_k \boldsymbol{\tau}_k$ with $a_j \neq 0$, then $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{j-1}, \boldsymbol{\eta}, \boldsymbol{\tau}_{j+1}, \ldots, \boldsymbol{\tau}_K$ are linearly independent.*

2. *Let R_k be the residue of Q_k with respect to Q_j for all $k \in [K]\backslash\{j\}$. Then, $R_k = \boldsymbol{\eta}_k^T \boldsymbol{P}$ where $\boldsymbol{\eta}_k \in \Delta_L$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{j-1}, \boldsymbol{\tau}_j, \boldsymbol{\eta}_{j+1}, \ldots, \boldsymbol{\eta}_K$ are linearly independent.*

3. *Let $\boldsymbol{\tau}^* \in \mathrm{conv}(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_k)^\circ$ and $\boldsymbol{\eta}_i \in \mathrm{conv}(\boldsymbol{\tau}_i, \boldsymbol{\tau}^*)^\circ$ for $i \in [k]$ where $k \leqslant K$. Then,*

$$\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k, \boldsymbol{\tau}_{k+1}, \ldots, \boldsymbol{\tau}_K$$

are linearly independent.

Proof We use the multilinear expansion and usual properties of determinants.

1. Viewing each $\boldsymbol{\tau}_i$ as a column vector,
$$\det(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{j-1}, \sum_k a_k \boldsymbol{\tau}_k, \boldsymbol{\tau}_{j+1}, \ldots, \boldsymbol{\tau}_K) = a_j \det(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K) \neq 0.$$

2. Linear independence of $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K$ implies that the $Q_1, \ldots, Q_K$ are distinct. Hence, by Proposition 2, we can write $\boldsymbol{\eta}_k = (1-\alpha_k)\boldsymbol{\tau}_j + \alpha_k \boldsymbol{\tau}_k$ where $\alpha_k \neq 0 \, \forall k \neq j$. Then, it holds that
$$\det(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{j-1}, \boldsymbol{\tau}_j, \boldsymbol{\eta}_{j+1}, \ldots, \boldsymbol{\eta}_K) = \left(\prod_{i \neq j} \alpha_i \right) \det(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K) \neq 0.$$

3. Since $\boldsymbol{\eta}_j = (1-\alpha_j)\boldsymbol{\tau}^* + \alpha_j \boldsymbol{\tau}_j$ where $\alpha_j \in (0,1)$ for all $j \leqslant k$, and $\boldsymbol{\tau}^* = \sum_i \beta_i \boldsymbol{\tau}_i$, it holds
$$\det(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{k-1}, \boldsymbol{\tau}_k, \ldots, \boldsymbol{\tau}_K) = \left(1 + \sum_{j=1}^{k} \frac{(1-\alpha_j)}{\alpha_j} \beta_j \right) \left(\prod_{i=1}^{k} \alpha_i \right) \det(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K) \neq 0.$$

■

Lemma 6 gives a condition on the mixture proportions under which the multi-sample residue is unique. Lemma 2 in Blanchard and Scott (2014) is very similar and is proved in a very similar way. We give a useful generalization here that reproduces many of the same details.

Lemma 6 *Let $l, k \in [L]$. Let $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L \in \Delta_L$ be linearly independent. We have that condition 1 implies condition 2 and condition 2 implies condition 3.*

1. *There exists a decomposition*
$$\boldsymbol{\tau}_l = \kappa \boldsymbol{e}_k + (1-\kappa)\boldsymbol{\tau}_l'$$
where $\kappa > 0$ and $\boldsymbol{\tau}_l' \in \operatorname{conv}(\{\boldsymbol{\tau}_j : j \neq l\})$. Further, for every $\boldsymbol{e}_i$ such that $i \neq k$, there exists a decomposition
$$\boldsymbol{e}_i = \sum_{j=1}^{L} a_j \boldsymbol{\tau}_j$$
such that $a_l < \frac{1}{\kappa}$.

2. *Let*
$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{\tau}_1^T \\ \vdots \\ \boldsymbol{\tau}_L^T \end{pmatrix};$$
the matrix $\boldsymbol{T}$ is invertible and $\boldsymbol{T}^{-1}$ is such that $(\boldsymbol{T}^{-1})_{l,k} > 0$ and $(\boldsymbol{T}^{-1})_{l,i} \leqslant 0$ for $i \neq k$ and $(\boldsymbol{T}^{-1})_{l,k} > (\boldsymbol{T}^{-1})_{j,k}$ for $j \neq l$. In words, the (l,k)th entry in $\boldsymbol{T}^{-1}$ is positive, every other entry in the lth row of $\boldsymbol{T}^{-1}$ is nonpositive and every other entry in the kth column of $\boldsymbol{T}^{-1}$ is strictly less than the (l,k)th entry. [2]

2. $(\boldsymbol{T}^{-1})_{i,j}$ is the $i \times j$ entry in the matrix $\boldsymbol{T}^{-1}$.

3. The residue of $\boldsymbol{\tau}_l$ with respect to $\{\boldsymbol{\tau}_j, j \neq l\}$ is $\boldsymbol{e}_k$.

Proof Without loss of generality, let $l = 1$ and $k = 2$. By relabeling the vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_L$, we can assume without loss of generality that $k = 1$. First, we show that condition *1* implies condition *2*. Suppose that condition *1* holds. Then, there exists $\kappa > 0$ such that

$$\boldsymbol{\tau}_1 = \kappa \boldsymbol{e}_1 + (1-\kappa)\sum_{i=2}^{L} \mu_i \boldsymbol{\tau}_i$$

with $\mu_i \geqslant 0$ for $i \in [L]\backslash\{1\}$. Then,

$$\boldsymbol{e}_1 = \frac{1}{\kappa}(\boldsymbol{\tau}_1 - \sum_{i \geqslant 2}(1-\kappa)\mu_i \boldsymbol{\tau}_i).$$

Hence, the first row of $\boldsymbol{T}^{-1}$ is given by $\frac{1}{\kappa}(1, -(1-\kappa)\mu_2, \cdots, -(1-\kappa)\mu_L)$. This shows that the first row is such that $(\boldsymbol{T}^{-1})_{1,1} > 0$ and $(\boldsymbol{T}^{-1})_{1,i} \leqslant 0$ for $i \neq 1$.

Consider $\boldsymbol{e}_i$ such that $i \neq 1$. Then, we have the relation: $\boldsymbol{e}_i = \sum_{j=1}^{L} a_j \boldsymbol{\tau}_j$, which gives the ith row of $\boldsymbol{T}^{-1}$. By assumption, $a_1 < \frac{1}{\kappa}$, so the $(i,1)$th entry is strictly less than the $(1,1)$th entry. Hence, *2* follows.

Now, we prove that condition *2* implies condition *3*. Suppose condition *2* is true. Consider the optimization problem

$$\max_{\boldsymbol{\nu}, \boldsymbol{\gamma}} \sum_{i=2}^{L} \nu_i \ \ s.t. \ \ \boldsymbol{\tau}_1 = (1 - \sum_{i \geqslant 2} \nu_i)\boldsymbol{\gamma} + \sum_{i=2}^{L} \nu_i \boldsymbol{\tau}_i$$

over $\boldsymbol{\gamma} \in \Delta_L$ and $\boldsymbol{\nu} = (\nu_2, \cdots, \nu_L) \in C_{L-1} = \{(\nu_2, \cdots, \nu_L) : \nu_i \geqslant 0; \sum_{i=2}^{L} \nu \leqslant 1\}$.

By the same argument given in the proof of Lemma 2 of Blanchard and Scott (2014), this optimization problem is equivalent to the program

$$\max_{\boldsymbol{\gamma} \in \Delta_L} \boldsymbol{e}_1^T (\boldsymbol{T}^T)^{-1} \boldsymbol{\gamma} \ \ s.t. \ \ \boldsymbol{\nu}((\boldsymbol{T}^T)^{-1}\boldsymbol{\gamma}) \in C_{L-1}$$

where $\boldsymbol{\nu}(\boldsymbol{\eta}) \coloneqq \eta_1^{-1}(-\eta_2, \cdots, -\eta_L)$. The above objective is of the form $\boldsymbol{a}^T\boldsymbol{\gamma}$ where $\boldsymbol{a}$ is the first column of $\boldsymbol{T}^{-1}$. Since $l = 1$, by assumption, for every $i \neq 1$, $\boldsymbol{T}_{1,1}^{-1} > \boldsymbol{T}_{i,1}^{-1}$. Therefore, the unconstrained maximum over $\boldsymbol{\gamma} \in \Delta_L$ is attained uniquely by $\boldsymbol{\gamma} = \boldsymbol{e}_1$. Notice that $(\boldsymbol{T}^T)^{-1}\boldsymbol{e}_1$ is the first row of $\boldsymbol{T}^{-1}$. Denote this vector $\boldsymbol{b} = (b_1, \cdots, b_L)$. We show that $\boldsymbol{\nu}(\boldsymbol{b}) = b_1^{-1}(-b_2, \cdots, -b_L) \in C_{L-1}$. By assumption, $\boldsymbol{b}$ has its first coordinate positive and the other coordinates are nonpositive. Therefore, all of the components of $\boldsymbol{\nu}(\boldsymbol{b})$ are nonnegative. Furthermore, the sum of the components of $\boldsymbol{\nu}(\boldsymbol{b})$ is

$$\sum_{i=2}^{L} \frac{-b_i}{b_1} = 1 - \frac{\sum_{i=1}^{L} b_i}{b_1} = 1 - \frac{1}{b_1} \leqslant 1.$$

The last equality follows because the rows of $\boldsymbol{T}^{-1}$ sum to 1 since $\boldsymbol{T}$ is a stochastic matrix. Then, we have $\boldsymbol{\nu}((\boldsymbol{T}^T)^{-1}\boldsymbol{e}_1) \in C_{L-1}$. Consequently, the unique maximum of the optimization problem is attained for $\boldsymbol{\gamma} = \boldsymbol{e}_1$. This establishes *3*. ■

D.3.2. The FaceTest Algorithm

Next, we consider the main subroutine in the Demix algorithm: the FaceTest algorithm (see Algorithm 6). Proposition 8 establishes that FaceTest$(Q_1, \ldots, Q_K)$ returns 1 if and only if $Q_1, \ldots, Q_K$ are in the relative interior of the same face of the simplex.

Proposition 8 *Let $Q_j = \boldsymbol{\eta}_j^T \boldsymbol{P}$ for $\boldsymbol{\eta}_j \in \Delta_K$ and all $j \in [K]$. Let $P_1, \ldots, P_K$ be jointly irreducible, $Q_1, \ldots, Q_K \in \text{conv}(P_1, \ldots, P_K)$ be distinct, and for each $i \in [K]$, let $\boldsymbol{\eta}_i$ lie in the relative interior of one of the faces of Δ_K. FaceTest($Q_1, \ldots, Q_K$) returns 1 if and only if $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ lie in the relative interior of the same face of Δ_K.*

Proof Suppose that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ lie on the relative interior of the same face of Δ_K. Then, $\mathcal{N}(Q_1) = \ldots = \mathcal{N}(Q_K)$. By statement 3 of Lemma 2, $\kappa^*(Q_i \,|Q_j) > 0$ for all $i \neq j$. Hence, FaceTest$(Q_1, \ldots, Q_K)$ returns 1.

Suppose that $Q_1, \ldots, Q_K$ do not all lie on the relative interior of the same face. Then, there exists Q_i, Q_j $(i \neq j)$ that do not lie on the relative interior of the same face. Without loss of generality, suppose that $\mathcal{N}(Q_j) \nsubseteq \mathcal{N}(Q_i)$. Then, by statement 3 of Lemma 2, $\kappa^*(Q_i \,|\, Q_j) = 0$. Hence, FaceTest$(Q_1, \ldots, Q_K)$ returns 0. ■

D.3.3. The Demix Algorithm

Proof [Proof of Theorem 2]

Let $K \leqslant L$, $\boldsymbol{\gamma}_i \in \Delta_L$ for all $i \in [K]$, $S_i = \boldsymbol{\gamma}_i^T \boldsymbol{P}$ for all $i \in [K]$, and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1^T \\ \vdots \\ \boldsymbol{\gamma}_K^T \end{pmatrix}.$$

We claim that for any $\{i_1, \ldots, i_K\} \subset [L]$ and $\{S_1, \ldots, S_K\} \subset \text{conv}(P_{i_1}, \ldots, P_{i_K})$, if $P_1, \ldots, P_L$ are jointly irreducible, and $\boldsymbol{\Gamma}$ has full row rank, then w.p. 1 Demix$(S_1, \ldots, S_K)$ returns a permutation of $(P_{i_1}, \ldots, P_{i_K})$. If the claim holds, then setting $K = L$ and putting $\tilde{P}_i = S_i$ yields the result. We prove the claim by induction on K.

Consider the base case: $K = 2$. Suppose that $\{S_1, S_2\} \subset \text{conv}(P_1, P_2)$ (the other cases are similar). Note that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_2$ by linear independence of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. Either $\boldsymbol{\gamma}_1 \in \text{conv}(\boldsymbol{e}_1, \boldsymbol{\gamma}_2)$ or $\boldsymbol{\gamma}_1 \in \text{conv}(\boldsymbol{e}_2, \boldsymbol{\gamma}_2)$. Suppose $\boldsymbol{\gamma}_1 \in \text{conv}(\boldsymbol{e}_1, \boldsymbol{\gamma}_2)$. Condition *2* of Lemma 1 is satisfied so that $\boldsymbol{e}_1$ is the residue of $\boldsymbol{\gamma}_1$ with respect to $\boldsymbol{\gamma}_2$ and $\boldsymbol{e}_2$ is the residue of $\boldsymbol{\gamma}_2$ with respect to $\boldsymbol{\gamma}_1$. Thus, by statement *3* of Proposition 3, P_1 is the residue of S_1 with respect to S_2 and P_2 is the residue of S_2 with respect to S_1. If $\boldsymbol{\gamma}_1 \in \text{conv}(\boldsymbol{e}_2, \boldsymbol{\gamma}_2)$, then similar reasoning establishes that P_2 is the residue of S_1 with respect to S_2 and P_1 is the residue of S_2 with respect to S_1. Thus, the base case follows.

Suppose $L \geqslant K > 2$. The inductive hypothesis is:

Inductive Hypothesis: for any $\{i_1, \ldots, i_{K-1}\} \subset [L]$ and $\{S_1, \ldots, S_{K-1}\} \subset \text{conv}(P_{i_1}, \ldots, P_{i_{K-1}})$, if $P_1, \ldots, P_L$ are jointly irreducible and $\boldsymbol{\Gamma}$ has full row rank, then w.p. 1 Demix$(S_1, \ldots, S_{K-1})$ returns a permutation of $(P_{i_1}, \ldots, P_{i_{K-1}})$.

Suppose that $\{S_1, \ldots, S_K\} \subset \text{conv}(P_1, \ldots, P_K)$ (the other cases are similar). Set $\Xi = \text{conv}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K)$. With probability 1, $Q \in \text{conv}(S_2, \ldots, S_K)^\circ$. We can write $Q = \boldsymbol{\eta}^T \boldsymbol{P}$ where $\boldsymbol{\eta}$ is a uniformly distributed random vector in $\text{conv}(\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)$. Let R be the residue of Q with respect to S_1. By statement *3* of Proposition 3, we can write $R = \boldsymbol{\lambda}^T \boldsymbol{P}$ where $\boldsymbol{\lambda}$ is the residue of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\gamma}_1$. By statement *2* of Lemma 2, $\boldsymbol{\lambda} \in \partial\Xi$.

Step 1: We claim that with probability 1, there is $l \in [K]$ such that $\boldsymbol{\lambda} \in \text{conv}(\{\boldsymbol{e}_j : j \in [K] \backslash \{l\}\})^\circ$. Let $B_{i,j} = \text{conv}(\{\boldsymbol{\gamma}_1\} \cup \{\boldsymbol{e}_k : k \in [K] \backslash \{i, j\}\})$ where $i, j \in [K]$ and $i \neq j$ and let $C = \text{conv}(\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)$. First, we argue that $C \cap B_{i,j}$ has affine dimension at most $K - 3$.[3] Since $\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K$ are linearly independent, C has affine dimension $K - 2$. Since $\{\boldsymbol{e}_k : k \in [K] \backslash \{i, j\}\}$ are linearly independent, $B_{i,j}$ has affine dimension $K-2$ or $K-3$. If $B_{i,j}$ has affine dimension $K-3$, then $C \cap B_{i,j}$ has affine dimension at most $K - 3$. So, suppose that $B_{i,j}$ has affine dimension $K - 2$. If $C \cap B_{i,j}$ has affine dimension $K - 2$, then $\text{aff}\, C = \text{aff}\, B_{i,j}$. Then, in particular, $\boldsymbol{\gamma}_1 \in \text{aff}\, C$. But, this contradicts the linear independence of $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$. Therefore, $C \cap B_{i,j}$ has affine dimension at most $K - 3$.

Because C has affine dimension $K-2$ and $\boldsymbol{\eta}$ is a uniformly distributed random vector in C, with probability 1, $\boldsymbol{\eta} \notin \cup_{i,j \in [K], i \neq j} B_{i,j}$. Since $\boldsymbol{\gamma}_1 \in B_{i,j}$ for all $i, j \in [K]$ and $\boldsymbol{\eta} \in \text{conv}(\boldsymbol{\lambda}, \boldsymbol{\gamma}_1)$ by definition, the convexity of $B_{i,j}$ implies that $\boldsymbol{\lambda} \notin \cup_{i,j \in [K], i \neq j} B_{i,j}$. Since $\boldsymbol{\lambda} \in \partial\Xi$, the claim follows.

Step 2: Let $R_i^{(n)}$ be the residue of $m_{\frac{n-1}{n}}(S_i, Q)$ with respect to S_1. We claim that there is some finite integer $N \geqslant 2$ such that for all $n \geqslant N$,

$$\text{FaceTest}(R_2^{(n)}, \ldots, R_K^{(n)})$$

returns 1. By Proposition 8, this is equivalent to the statement that there exists $N \geqslant 2$ such that for all $n \geqslant N$, the mixture proportions of $R_2^{(n)}, \ldots, R_K^{(n)}$ are on the relative interior of the same face. Let $m_{\frac{n-1}{n}}(S_i, Q) = (\boldsymbol{\tau}_i^{(n)})^T \boldsymbol{P}$ for $i \in [K] \backslash \{1\}$; note that $\boldsymbol{\tau}_i^{(n)} = \frac{1}{n}\boldsymbol{\gamma}_i + \frac{n-1}{n}\boldsymbol{\eta}$ and, consequently, $\boldsymbol{\tau}_i^{(n)} \in \Xi$. Since $\boldsymbol{\eta} \in \text{conv}(\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)^\circ$ with probability 1, $\boldsymbol{\tau}_i^{(n)} \in \text{conv}(\boldsymbol{\gamma}_i, \boldsymbol{\eta})^\circ$ for all $i \in [K] \backslash \{1\}$ and $n \in \mathbb{N}$, and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ are linearly independent, it follows that for all $n \in \mathbb{N}$ with probability 1, $\boldsymbol{\gamma}_1, \boldsymbol{\tau}_2^{(n)}, \ldots, \boldsymbol{\tau}_K^{(n)}$ are linearly independent by statement *3* in Lemma 5. Fix $i \in [K] \backslash \{1\}$. It suffices to show that there is large enough N such that for $n \geqslant N$, $\text{Residue}(m_{\frac{n-1}{n}}(S_i, Q) \,|\, S_1) = R_i^{(n)}$ is on the same face as R. Let $R_i^{(n)} = (\boldsymbol{\mu}_i^{(n)})^T \boldsymbol{P}$; by statement *3* of Proposition 3, $\boldsymbol{\mu}_i^{(n)}$ is the residue of $\boldsymbol{\tau}_i^{(n)}$ with respect to $\boldsymbol{\gamma}_1$ and by statement 2 of Lemma 2 $\boldsymbol{\mu}_i^{(n)} \in \Xi$. It suffices to show that $\mathcal{N}(\boldsymbol{\mu}_i^{(n)}) = \mathcal{N}(\boldsymbol{\lambda})$, i.e., every $\boldsymbol{\mu}_i^{(n)}$ is on the same face as $\boldsymbol{\lambda}$. As $n \longrightarrow \infty$, $\boldsymbol{\tau}_i^{(n)} = (1 - \frac{n-1}{n})\boldsymbol{\gamma}_i + \frac{n-1}{n}\boldsymbol{\eta} \longrightarrow \boldsymbol{\eta}$, hence by statement *2* in Lemma 4, $\left\|\boldsymbol{\mu}_i^{(n)} - \boldsymbol{\lambda}\right\| \longrightarrow 0$. Since with probability 1, $\boldsymbol{\lambda} \in \text{conv}(\{\boldsymbol{e}_j : j \in [K] \backslash \{l\}\})^\circ$ for some l (step 1), it follows that for large enough n, $\boldsymbol{\mu}_i^{(n)} \in \text{conv}(\{\boldsymbol{e}_j : j \in [K] \backslash \{l\}\})^\circ$.

3. Note that if $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^L$ are linearly independent and $n \leqslant L$, then $\text{aff}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ has affine dimension $n - 1$.

Algorithm 15 NonSquareDemix($\tilde{P}_1, \ldots, \tilde{P}_M$)

1: $R_1, \ldots, R_L \longleftarrow$ independently uniformly distributed elements in $\operatorname{conv}(\tilde{P}_1, \ldots, \tilde{P}_M)$
2: $(Q_1, \ldots, Q_L)^T \longleftarrow \text{Demix}(R_1, \ldots, R_L)$
3: **return** $(Q_1, \ldots, Q_L)^T$

Step 3: Assume that n is sufficiently large such that $R_2^{(n)}, \ldots, R_K^{(n)}$ are on the same face. The algorithm recurses on $R_2^{(n)}, \ldots, R_K^{(n)}$. Since $\boldsymbol{\gamma}_1, \boldsymbol{\tau}_2^{(n)}, \ldots, \boldsymbol{\tau}_K^{(n)}$ are linearly independent, it follows by statement *2* in Lemma 5 that $\boldsymbol{\mu}_2^{(n)}, \ldots, \boldsymbol{\mu}_K^{(n)}$ are linearly independent. Suppose wlog that $\{R_2^{(n)}, \ldots, R_K^{(n)}\} \subset \{P_1, \ldots, P_{K-1}\}$. Then, by the inductive hypothesis, if $(Q_1, \ldots, Q_{K-1}) \longleftarrow \text{Demix}(R_2^{(n)}, \ldots, R_K^{(n)})$, then $(Q_1, \ldots, Q_{K-1})$ is a permutation of $(P_1, \ldots, P_{K-1})$. Note that $\frac{1}{K}\sum_{i=1}^K S_i \in \operatorname{conv}(P_1, \ldots, P_K)^\circ$ since $\boldsymbol{\Gamma}$ has full rank by assumption.

Write $Q_i = \boldsymbol{\rho}_i^T \boldsymbol{P}$ for $i \in [K]$. Then, there exists of a permutation $\sigma : [K-1] \longrightarrow [K-1]$ such that $\boldsymbol{\rho}_i = \boldsymbol{e}_{\sigma(i)}$. Since $\boldsymbol{\rho}_K \in \Xi^\circ$ and $\boldsymbol{\rho}_i = \boldsymbol{e}_{\sigma(i)}$ for $i \leqslant K-1$, the conditions in statement *1* of Lemma 6 are satisfied. Therefore, by Lemma 6, the residue of $\boldsymbol{\rho}_K$ with respect to $\{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{K-1}\}$ is $\boldsymbol{e}_K$. Then, by statement *3* of Proposition 3, the residue of Q_K with respect to $\{Q_1, \ldots, Q_{K-1}\}$ is P_K. This completes the inductive step.

■

D.3.4. The Non-Square Demix Algorithm

Now, we examine the non-square case of the demixing problem ($M > L$). Note that knowledge of L is needed since one must resample exactly L distributions in order to run the square Demix algorithm.

Corollary 1 *Suppose $M > L$. Let $P_1, \ldots, P_L$ be jointly irreducible and $\boldsymbol{\Pi}$ have full rank. Then, with probability 1, NonSquareDemix($\tilde{P}_1, \ldots, \tilde{P}_M$) returns $(Q_1, \ldots, Q_L)$ such that $(Q_1, \ldots, Q_L)$ is a permutation of $(P_1, \ldots, P_L)$.*

Proof We can write $R_i = \boldsymbol{\tau}_i^T \boldsymbol{P}$ where $\boldsymbol{\tau}_i \in \Delta_L$ and $i = 1, \ldots, L$. $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L$ are drawn uniformly independently from a set with positive $(L-1)$-dimensional Lebesgue measure since $\boldsymbol{\Pi}$ has full rank by hypothesis. By Lemma 3, $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L$ are linearly independent with probability 1. Then, by Theorem 2, with probability 1, Demix($R_1, \ldots, R_L$) returns a permutation of $(P_1, \ldots, P_L)$. ■

D.4. Classification with Partial Labels

In this section, we present our identification result for classification with partial labels, i.e., Theorem 3. To begin, in Section D.4.1, we prove an important lemma for the main subroutine of the algorithm PartialLabel: VertexTest (algorithm 9). Second, in Section D.4.2, we present the proof of Theorem 3.

D.4.1. VertexTest Algorithm

Lemma 7 establishes that the VertexTest algorithm determines whether one vector of distributions is a permutation of another vector of distributions.

Lemma 7 *Let $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_L \in \Delta_L$ and $Q_i = \boldsymbol{\eta}_i^T \boldsymbol{P}$ for $i \in [L]$ and $\boldsymbol{Q} = (Q_1, \dots, Q_L)^T$. Suppose that $P_1, \dots, P_L$ are jointly irreducible, $\boldsymbol{\Pi}$ has full column rank, and the columns of $\boldsymbol{\Pi}^+$ are unique. Then, VertexTest$(\boldsymbol{\Pi}^+, \tilde{\boldsymbol{P}}, \boldsymbol{Q})$ returns $(1, \boldsymbol{C}^T)$ with $\boldsymbol{C}$ a permutation matrix if and only if $\boldsymbol{Q}$ is a permutation of $\boldsymbol{P}$. Further, if VertexTest$(\boldsymbol{\Pi}^+, \tilde{\boldsymbol{P}}, \boldsymbol{Q})$ returns $(1, \boldsymbol{C}^T)$, then $\boldsymbol{C}^T \boldsymbol{Q} = \boldsymbol{P}$.*

Proof If $\boldsymbol{Q} = (Q_1, \dots, Q_L)^T$ is such that $\boldsymbol{D}^T \boldsymbol{Q} = \boldsymbol{P}$ where $\boldsymbol{D}$ is a permutation matrix, then it is clear that VertexTest$(\boldsymbol{\Pi}^+, (\tilde{P}_1, \dots, \tilde{P}_M)^T, (Q_1, \dots, Q_L)^T)$ returns $(1, \boldsymbol{C}^T)$ for some permutation matrix $\boldsymbol{C}$ since the entries of $\boldsymbol{\Pi}^+$ are $\boldsymbol{\Pi}^+_{i,j} = \mathbf{1}_{\{\kappa^*(\tilde{P}_i \,|\, P_j) > 0\}}$. Since $\boldsymbol{D}^T \boldsymbol{Q} = \boldsymbol{P}$, clearly, $\boldsymbol{Z}\boldsymbol{D} = \boldsymbol{\Pi}^+$. But since the columns of $\boldsymbol{\Pi}^+$ are unique, there is a unique permutation of the columns of $\boldsymbol{Z}$ to obtain the columns of $\boldsymbol{\Pi}^+$. Therefore, $\boldsymbol{D} = \boldsymbol{C}$.

Consider the "only if" direction. We use the notation from Algorithm 9. Suppose Algorithm 9 has returned $(1, \boldsymbol{C}^T)$ where $\boldsymbol{C}$ is a permutation matrix. W.l.o.g. (reordering the Q_i) we can assume that $\boldsymbol{C}$ is the identity and thus $\boldsymbol{Z} = \boldsymbol{\Pi}^+$.

In the sequel denote $\phi(x) := \mathbf{1}_{\{x>0\}}$ and $\phi(\boldsymbol{M})$ the entry-wise application of ϕ to the matrix or vector $\boldsymbol{M}$. We denote $\boldsymbol{v} \preceq \boldsymbol{w}$ when all entries of $\boldsymbol{v}$ are less than or equal to the corresponding entries of $\boldsymbol{w}$ (where $\boldsymbol{v}$ and $\boldsymbol{w}$ are vectors). This is a partial order, which will be used only for $0-1$ vectors below (essentially to denote support inclusion). W.l.o.g. (reordering the P_i) we can assume that the columns of $\boldsymbol{\Pi}^+$ are reordered in some sequence compatible with $\preceq$ in decreasing order, i.e. such that if $\boldsymbol{\Pi}^+_{:,j} \preceq \boldsymbol{\Pi}^+_{:,i}$, then $i \leqslant j$.

Introduce the following additional notation: let $\boldsymbol{\Lambda}$ be the matrix with rows $\boldsymbol{\Lambda}_{i,:} = \phi(\boldsymbol{\eta}_i^T)$. Observe that by statement 3 of Lemma 2, for any i, j, k, $\kappa^*(\tilde{P}_i | Q_j) > 0$ and $\kappa^*(Q_j | P_k) > 0$ implies $\kappa^*(\tilde{P}_i | P_k) > 0$. Note that we can write $\boldsymbol{\Lambda}_{j,k} = \mathbf{1}_{\{\kappa^*(Q_j \,|\, P_k) > 0\}}$ and $\boldsymbol{\Pi}^+_{j,k} = \mathbf{1}_{\{\kappa^*(\tilde{P}_j \,|\, P_k) > 0\}}$. Thus, we must have $\phi(\boldsymbol{Z}\boldsymbol{\Lambda}) \preceq \boldsymbol{\Pi}^+$.

We now argue that this implies that $\boldsymbol{\Lambda}$ is sub-diagonal, i.e., $\boldsymbol{\Lambda}_{ij} = 0$ for $i < j$. Let $i < j$. If $\boldsymbol{\Lambda}_{ij} > 0$, then $\boldsymbol{Z}_{:,i} \preceq \boldsymbol{\Pi}^+_{:,j}$ by the above relation. Since $\boldsymbol{Z} = \boldsymbol{\Pi}^+$, this implies $\boldsymbol{\Pi}^+_{:,i} \preceq \boldsymbol{\Pi}^+_{:,j}$, which implies $j \leqslant i$ by the assumed ordering of the columns of $\boldsymbol{\Pi}^+$, a contradiction. Hence $\boldsymbol{\Lambda}_{ij} = 0$ for $i < j$.

Now, since the matrix $\boldsymbol{Y}$ (line 1 of Algorithm 9) is diagonal, Statement 3 of Lemma 1 gives that for any $i \neq j$ we have $\boldsymbol{\Lambda}_{i,:} \not\preceq \boldsymbol{\Lambda}_{j,:}$. One can conclude by a straightforward recursion that since $\boldsymbol{\Lambda}$ is sub-diagonal, this implies that $\boldsymbol{\Lambda}$ is in fact diagonal. Start with the first row $\boldsymbol{\Lambda}_{1,:}$ which must be $(1, 0, \dots, 0)^T$ (by sub-diagonality). Since $\boldsymbol{\Lambda}_{1,:} \not\preceq \boldsymbol{\Lambda}_{j,:}$ for $j > 1$, this implies the first column $\boldsymbol{\Lambda}_{:,1}$ is also $(1, 0, \dots, 0)$. The subsequent columns/rows are handled in the same way.

Hence $\boldsymbol{\Lambda}$ is the identity, which implies that $\boldsymbol{Q} = \boldsymbol{P}$. ■

D.4.2. Proof of Theorem 3

Proof We adopt the notation from the description of Algorithm 7 with the exception that we make explicit the dependence on k by writing $W_i^{(k)}$ instead of W_i and $\bar{Q}_i^{(k)}$ instead of

$\bar{Q}_i$. We show that there is a K such that for all $k \geqslant K$, $(W_1^{(k)}, \ldots, W_L^{(k)})^T$ is a permutation of $(P_1, \ldots, P_L)^T$. Then, the result will follow from Lemma 7.

Let $Q_i = \boldsymbol{\tau}_i^T \boldsymbol{P}$, $\bar{Q}_i^{(k)} = \bar{\boldsymbol{\tau}}_i^{(k)^T} \boldsymbol{P}$, and $W_i^{(k)} = \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{P}$. Further, let $0 \leqslant n < L$, $\{i_1, \ldots, i_n\} \subset [L]$, $l \neq j \in [L]$, and define the following events wrt the randomness of $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L$:

$$
\begin{aligned}
E_{i_1,\ldots,i_n} &= \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_n}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L \text{ are linearly independent}\} \\
E &= \cap_{\{i_1,\ldots,i_n\} \subset [L], 0 \leqslant n < L} E_{i_1,\ldots,i_n} \\
F_{l,j} &= \{\boldsymbol{e}_l - \boldsymbol{e}_j, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L \text{ are linearly independent}\} \\
F &= \cap_{l \neq j \in [L]} F_{l,j} \\
G^{l,j}_{i_1,\ldots,i_{n-1}} &= \{\boldsymbol{e}_l - \boldsymbol{e}_j, \boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L \text{ are linearly independent}\} \\
G &= \cap_{\{i_1,\ldots,i_{n-1}\} \subset [L], 0 \leqslant n < L, l \neq j \in [L] \setminus \{i_1,\ldots,i_{n-1}\}} G^{l,j}_{i_1,\ldots,i_{n-1}}.
\end{aligned}
$$

By Lemma 3, for any $0 \leqslant n < L$, $\{i_1, \ldots, i_n\} \subset [L]$, the event $E_{i_1,\ldots,i_n}$ occurs with probability 1. Similarly, by Lemma 3, for any $l \neq j \in [L]$, the event $F_{l,j}$ occurs with probability 1. Finally, by Lemma 3, for any $0 \leqslant n < L$, $\{i_1, \ldots, i_{n-1}\} \subset [L]$ and any $l \neq j \in [L] \setminus \{i_1, \ldots, i_{n-1}\}$, the event $G^{l,j}_{i_1,\ldots,i_{n-1}}$ occurs with probability 1. Hence, the event $E \cap F \cap G$ occurs with probability 1. For the remainder of the proof, assume event $E \cap F \cap G$ occurs.

We prove the claim inductively. We show that for all $n \leqslant L$ there exists K_n such that if $k \geqslant K_n$, then $W_1^{(k)}, \ldots, W_n^{(k)}$ are distinct base distributions.

Base Case: $n = 1$**.** We will apply Lemma 6. By event E, $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L$ are linearly independent. Therefore, $\text{aff}(\boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L)$ gives a hyperplane with an associated open halfspace $\boldsymbol{H}$ that contains $\boldsymbol{\tau}_1$ and at least one $\boldsymbol{e}_j$. Inspection of Line 3 of Algorithm 8 shows that $\bar{\boldsymbol{\tau}}_1^{(k)}$ is simply the average of $\boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L$ and does not depend on k. Thus, there exists K_1 such that for all $k \geqslant K_1$, if $\boldsymbol{e}_j \in \boldsymbol{H}$, then $\boldsymbol{\lambda}_k \coloneqq \frac{1}{k}\boldsymbol{\tau}_1 + \frac{k-1}{k}\bar{\boldsymbol{\tau}}_1^{(k)} \in \text{conv}(\boldsymbol{e}_j, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L)^\circ$. Fix $k \geqslant K_1$. Then, by event E, for all $\boldsymbol{e}_j \in \boldsymbol{H}$, there exists a unique $\kappa_j > 0$ and unique $a_{j,2}, \ldots, a_{j,L}$ such that

$$
\begin{aligned}
\boldsymbol{\lambda}_k &= \kappa_j \boldsymbol{e}_j + \sum_{i=2}^{L} a_{j,i} \boldsymbol{\tau}_i \\
&= \kappa_j \boldsymbol{e}_j + (1 - \kappa_j) \tilde{\boldsymbol{\tau}}_j
\end{aligned}
$$

where $\tilde{\boldsymbol{\tau}}_j \in \text{conv}(\boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L)$ is unique. We claim that for all $i \neq j$ and $\{\boldsymbol{e}_i, \boldsymbol{e}_j\} \subset \boldsymbol{H}$, $\kappa_i \neq \kappa_j$. Suppose to the contrary that there is $i \neq j$ such that $\{\boldsymbol{e}_i, \boldsymbol{e}_j\} \subset \boldsymbol{H}$ and $\kappa_i = \kappa_j = \kappa$. Then,

$$
\begin{aligned}
\boldsymbol{\lambda}_k &= \kappa \boldsymbol{e}_i + (1 - \kappa) \tilde{\boldsymbol{\tau}}_i \\
\boldsymbol{\lambda}_k &= \kappa \boldsymbol{e}_j + (1 - \kappa) \tilde{\boldsymbol{\tau}}_j.
\end{aligned}
$$

Then, $(1-\kappa)(\tilde{\boldsymbol{\tau}}_j - \tilde{\boldsymbol{\tau}}_i) - \kappa(\boldsymbol{e}_i - \boldsymbol{e}_j) = 0$, from which it follows that $\boldsymbol{e}_i - \boldsymbol{e}_j \in \text{span}(\boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L)$. But, by event F, $\boldsymbol{e}_i - \boldsymbol{e}_j, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L$ are linearly independent and, hence, we have a contradiction. Thus, the claim follows.

Consequently, there is a unique j that minimizes κ_j. Note that for all $\boldsymbol{e}_i \notin \boldsymbol{H}$, if we write $\boldsymbol{e}_i = \sum_{l \geqslant 2} a_l \boldsymbol{\tau}_l + a_1 \boldsymbol{\lambda}_k$, then $a_1 \leqslant 0$. Then, by Lemma 6, $\boldsymbol{e}_j$ is the residue of $\boldsymbol{\lambda}_k$ with

respect to $\boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L$. Therefore, by Proposition 3, MultiResidue$(\frac{1}{k}Q_1 + (1-\frac{1}{k})\bar{Q}_1 \,|\, \{Q_j\}_{j>1})$ is well-defined and if $W_1^{(k)} \longleftarrow$ MultiResidue$(\frac{1}{k}Q_1 + (1-\frac{1}{k})\bar{Q}_1 \,|\, \{Q_j\}_{j>1})$, $W_1^{(k)}$ is one of the base distributions. This establishes the base case.

The Inductive Step: The proof is similar to the base case. Suppose that there exists K_{n-1} such that for all $k \geqslant K_{n-1}$, $W_1^{(k)}, \ldots, W_{n-1}^{(k)}$ are distinct base distributions. Let $\{i_1, \ldots, i_{n-1}\} \subset [L]$ denote the indices of the base distributions that are equal to $W_1^{(k)}, \ldots, W_{n-1}^{(k)}$ under the inductive hypothesis. By the event E, $\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_n, \ldots, \boldsymbol{\tau}_L$ are linearly independent. Hence, aff$(\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L)$ gives a hyperplane with an associated open halfspace $\boldsymbol{H}_{i_1,\ldots,i_{n-1}}$ such that $\boldsymbol{\tau}_n \in \boldsymbol{H}_{i_1,\ldots,i_{n-1}}$. We claim that there is $\boldsymbol{e}_j \notin \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}\}$ such that $\boldsymbol{e}_j \in \boldsymbol{H}_{i_1,\ldots,i_{n-1}}$. Suppose not. Then, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_L \in \boldsymbol{H}^c_{i_1,\ldots,i_{n-1}}$ and $\boldsymbol{\tau}_n \in \boldsymbol{H}_{i_1,\ldots,i_{n-1}}$, which implies that $\boldsymbol{\tau}_n \notin \Delta_{L-1}$. This is a contradiction, so the claim follows.

Define

$$\boldsymbol{\lambda}_k^{(i_1,\ldots,i_{n-1})} \coloneqq \frac{1}{k}\boldsymbol{\tau}_n + [\frac{k-1}{k}]\frac{1}{L-1}(\sum_{s>n}\boldsymbol{\tau}_s + \sum_{s<n}\boldsymbol{e}_{i_s}).$$

There exists an integer $K_n^{(i_1,\ldots,i_{n-1})}$ such that if $k \geqslant K_n^{(i_1,\ldots,i_{n-1})}$, then for all $\boldsymbol{e}_j \in \boldsymbol{H}_{i_1,\ldots,i_{n-1}}$, $\boldsymbol{\lambda}_k^{(i_1,\ldots,i_{n-1})} \in \text{conv}(\boldsymbol{e}_j, \boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L)^\circ$. Set

$$K_n \coloneqq \max(\max_{\{i_1,\ldots,i_{n-1}\}\subset[L]}(K_n^{(i_1,\ldots,i_{n-1})}), K_{n-1}).$$

Fix $k \geqslant K_n$. Define

$$\begin{aligned}\boldsymbol{\lambda}_k &\coloneqq \frac{1}{k}\boldsymbol{\tau}_n + [\frac{k-1}{k}]\bar{\boldsymbol{\tau}}_n^{(k)} \\ &= \frac{1}{k}\boldsymbol{\tau}_n + [\frac{k-1}{k}]\frac{1}{L-1}(\sum_{s>n}\boldsymbol{\tau}_s + \sum_{s<n}\boldsymbol{\gamma}_s^{(k)}).\end{aligned}$$

By the inductive hypothesis, $k \geqslant K_{n-1}$, and Proposition 3, there exists $\{i_1, \ldots, i_{n-1}\} \subset [L]$ such that $\boldsymbol{\gamma}_j^{(k)} = \boldsymbol{e}_{i_j}$ for all $j \in [n-1]$. For the sake of abbreviation, let $\boldsymbol{H} = \boldsymbol{H}_{i_1,\ldots,i_{n-1}}$. Thus, $\boldsymbol{\tau}_n \in \boldsymbol{H}$ and there exists $\boldsymbol{e}_j \in \boldsymbol{H}$ such that $\boldsymbol{e}_j \notin \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}\}$. Hence, by our choice of K_n, for every $\boldsymbol{e}_j \in \boldsymbol{H}$

$$\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_k^{(i_1,\ldots,i_{n-1})} \in \text{conv}(\boldsymbol{e}_j, \boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L)^\circ.$$

By event E for all $\boldsymbol{e}_j \in \boldsymbol{H}$, there is a unique $\kappa_j > 0$ and unique $a_{j,1}, \ldots, a_{j,n-1}, a_{j,n+1}, \ldots, a_{j,L} > 0$ such that

$$\begin{aligned}\boldsymbol{\lambda}_k &= \kappa_j\boldsymbol{e}_j + \sum_{l<n} a_{j,i}\boldsymbol{e}_{i_l} + \sum_{l>n} a_{j,l}\boldsymbol{\tau}_l \\ &= \kappa_j\boldsymbol{e}_j + (1-\kappa_j)\tilde{\boldsymbol{\tau}}_j\end{aligned}$$

where $\tilde{\boldsymbol{\tau}}_j \in \text{conv}(\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L)$ is unique.

We claim that for all $l \neq j$ such that $\{\boldsymbol{e}_l, \boldsymbol{e}_j\} \subset \boldsymbol{H}$, $\kappa_l \neq \kappa_j$. Suppose to the contrary that there exists $l \neq j$ such that $\{\boldsymbol{e}_l, \boldsymbol{e}_j\} \subset \boldsymbol{H}$ and $\kappa_l = \kappa_j = \kappa$. Then,

$$\boldsymbol{\lambda}_k = \kappa \boldsymbol{e}_l + (1-\kappa)\tilde{\boldsymbol{\tau}}_i = \kappa \boldsymbol{e}_j + (1-\kappa)\tilde{\boldsymbol{\tau}}_j.$$

This implies that $\boldsymbol{e}_l - \boldsymbol{e}_j \in \text{span}(\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L)$. Observe that $\{\boldsymbol{e}_l, \boldsymbol{e}_j\} \subset \boldsymbol{H}$ implies that $\boldsymbol{e}_l \notin \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}\}$ and $\boldsymbol{e}_j \notin \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}\}$. Thus, event G implies that $\boldsymbol{e}_l - \boldsymbol{e}_j, \boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L$ are linearly independent. Therefore, we have a contradiction, establishing the claim.

Consequently, there is a unique j that minimizes κ_j. Note that for all $\boldsymbol{e}_l \notin \boldsymbol{H}$, if we write $\boldsymbol{e}_l = \sum_{m<n} a_m \boldsymbol{e}_{i_m} + \sum_{m>n} a_m \boldsymbol{\tau}_m + a_n \boldsymbol{\lambda}_k$, then $\boldsymbol{a}_n \leqslant 0$. Then, by Lemma 6, $\boldsymbol{e}_j$ is the residue of $\boldsymbol{\lambda}_k$ with respect to $\boldsymbol{\gamma}_1^{(k)}, \ldots, \boldsymbol{\gamma}_{n-1}^{(k)}, \boldsymbol{\tau}_{n+1}, \ldots, \boldsymbol{\tau}_L$. Therefore, by Proposition 3, MultiResidue$(\frac{1}{k}Q_n + (1-\frac{1}{k})\bar{Q}_n \,|\, \{Q_j\}_{j>n} \cup \{W_j^{(k)}\}_{j<n})$ is well-defined and if $W_n \longleftarrow$ MultiResidue$(\frac{1}{k}Q_n + (1-\frac{1}{k})\bar{Q}_n \,|\, \{Q_j\}_{j>n} \cup \{W_j^{(k)}\}_{j<n})$, $W_n^{(k)}$ is one of the base distributions. Since $\boldsymbol{e}_j \in \boldsymbol{H}$ implies that $\boldsymbol{e}_j \notin \{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{n-1}}\}$, it follows that $W_1^{(k)}, \ldots, W_n^{(k)}$ are distinct base distributions. This establishes the inductive step.

The result follows from applying Lemma 7. ■

Appendix E. Estimation

In this section, we present the estimation results of our paper. To begin, in Section E.1, we present the proof of sufficient conditions under which ResidueHat estimators converge uniformly in probability (Proposition 4). Second, in Section E.2, we prove our main estimation result for demixing mixed membership models (Theorem 5). Finally, in Section E.3, we prove our main estimation result for classification with partial labels (Theorem 6).

E.1. ResidueHat Results

Let $A_1, A_2, \ldots$ denote positive constants whose values may change from line to line. We introduce the following definitions.

Definition 11 *Let $\widehat{F}$ and $\widehat{H}$ be ResidueHat estimators of F and H, respectively, where $F \neq H$ and let $G \longleftarrow$ Residue$(F \,|\, H)$ and $\widehat{G} \longleftarrow$ ResidueHat$(\widehat{F} \,|\, \widehat{H})$. If $\widehat{G}$ is a ResidueHat estimator of order 0, we say its* distributional ancestors *are $\{F, H\}$ and define ancestors$(\widehat{G}) \coloneqq \{F, H\}$. If $\widehat{G}$ is a ResidueHat estimator of the kth order, we define its distributional ancestors to be ancestors$(\widehat{G})$ = ancestors$(\widehat{F}) \cup$ ancestors$(\widehat{H})$.*

The constants in our bounds depend on the distributional ancestors.

Definition 12 *We say that the distribution F satisfies the* support condition ***(SC)*** *with respect to H if there exists a distribution G and $\gamma \in [0, 1)$ such that* $\text{supp}(H) \nsubseteq \text{supp}(G)$ *and $F = (1-\gamma)G + \gamma H$.*

Definition 13 *If*

$$\sup_{E \in \mathcal{E}} |\widehat{F}(E) - F(E)| \xrightarrow{i.p.} 0$$

as $\boldsymbol{n} \longrightarrow \infty$*, we say that* $\widehat{F} \longrightarrow F$ *uniformly (or* $\widehat{F}$ *converges uniformly to* F*) with respect to* $\mathcal{E}$*.*

Definition 14 *Let* $\widehat{F}$ *be a ResidueHat estimator of a distribution* F*. We say that* $\widehat{F}$ *satisfies a Uniform Deviation Inequality* ***(UDI)*** *with respect to* $\mathcal{E}$ *if for all* $\epsilon > 0$*, there exist constants* $A_{1,\epsilon}, A_{2,\epsilon} > 0$ *and* $\boldsymbol{N}$ *depending on ancestors*$(\widehat{F})$ *such that if* $\boldsymbol{n} \geqslant \boldsymbol{N}$*, then for all* $E \in \mathcal{E}$

$$|\widehat{F}(E) - F(E)| < A_{1,\epsilon}\gamma_{\boldsymbol{n}} + \epsilon$$

with probability at least $1 - A_{2,\epsilon}\sum_{i\in[L]}\frac{1}{n_i}$

Henceforth, for the purposes of abbreviation, we will only say that a ResidueHat estimator satisfies a Uniform Deviation Inequality **(UDI)** and omit "with respect to $\mathcal{E}$" because the context makes this clear.

Definition 15 *Let* $\widehat{F}$ *and* $\widehat{H}$ *be ResidueHat estimators. We say that* $\widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$ *satisfies a Rate of Convergence* ***(RC)*** *with respect to* $\mathcal{E}$ *if for all* $\epsilon > 0$*, there exists constants* $A_{1,\epsilon}, A_{2,\epsilon} > 0$ *and* $\boldsymbol{N}$ *depending on ancestors*$(\widehat{F}) \cup$ *ancestors*$(\widehat{H})$ *such that for* $\boldsymbol{n} \geqslant \boldsymbol{N}$*,*

$$|\widehat{\kappa}(\widehat{F} \,|\, \widehat{H}) - \kappa^*(F \,|\, H)| \leqslant A_{1,\epsilon}\gamma_{\boldsymbol{n}} + \epsilon$$

with probability at least $1 - A_{2,\epsilon}\sum_{i\in[L]}\frac{1}{n_i}$*.*

Lemma 8 gives sufficient conditions under which F satisfies **(SC)** with respect to H.

Lemma 8 *Let* $P_1, \ldots, P_L$ *satisfy* ***(A″)*** *and let* $F, H \in \text{conv}(P_1, \ldots, P_L)$ *such that* $F \neq H$*. Then,* F *satisfies* ***(SC)*** *with respect to* H*.*

Proof Let $A = \arg\min(|B| : B \subseteq \{P_1, \ldots, P_L\}, F, H \in \text{conv}(B))$. Without loss of generality, suppose that $A = \{P_1, \ldots, P_K\}$. F either lies on the boundary of $\text{conv}(P_1, \ldots, P_K)$ or doesn't. If F lies on the boundary of $\text{conv}(P_1, \ldots, P_K)$, then $H \in \text{conv}(P_1, \ldots, P_K)^\circ$ by minimality of A. Then, we pick $G = F$ and $\gamma = 0$ to obtain $F = (1-\gamma)F + \gamma H$. Since $P_1, \ldots, P_L$ satisfy **(A″)**, $\text{supp}(H) \nsubseteq \text{supp}(F)$.

Now, suppose that $F \in \text{conv}(P_1, \ldots, P_K)^\circ$. Let $G \longleftarrow \text{Residue}(F \,|\, H)$; we can write $F = (1-\gamma)G + \gamma H$ for $\gamma \in [0, 1)$ since $F \neq H$. Then, by Statement *2* of Lemma 2 and statement *3* of Proposition 3, G is on the boundary of $\text{conv}(P_1, \ldots, P_K)$. Without loss of generality, suppose that $G \in \text{conv}(P_1, \ldots, P_{K-1})$. Since $F = (1-\gamma)G + \gamma H \in \text{conv}(P_1, \ldots, P_K)^\circ$, and $G \in \text{conv}(P_1, \ldots, P_{K-1})$, $H \notin \text{conv}(P_1, \ldots, P_{K-1})$. Since $P_1, \ldots, P_L$ satisfy **(A″)**, $\text{supp}(H) \nsubseteq \text{supp}(G)$. This completes the proof. ■

Lemma 9 gives sufficient conditions under which an estimator $\widehat{G}$ satisfies a **(UDI)**.

Lemma 9 *Let*

1. F *and* H *be distributions such that* $F \neq H$*,*
2. $G \longleftarrow$ *Residue*$(F \,|\, H)$*, and*
3. $\widehat{G} \longleftarrow$ *ResidueHat*$(\widehat{F} \,|\, \widehat{H})$*.*

*If $\widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$ satisfies a **(RC)**, $\widehat{H}$ satisfies a **(UDI)**, and $\widehat{F}$ satisfies a **(UDI)**, then $\widehat{G}$ satisfies a **(UDI)**.*

Proof For the sake of abbreviation, let $\widehat{\kappa} = \widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$, $\kappa^* = \kappa^*(F \,|\, H)$, $\widehat{\alpha} = \frac{1}{1-\widehat{\kappa}}$ and $\alpha^* = \frac{1}{1-\kappa^*}$. Let $\epsilon > 0$. We claim that there are constants $A_{1,\epsilon}, A_{2,\epsilon} > 0$ such that for sufficiently large $\boldsymbol{n}$,

$$\Pr(|\widehat{\alpha} - \alpha^*| < A_{1,\epsilon}\gamma_{\boldsymbol{n}} + \epsilon) \geqslant 1 - A_{2,\epsilon} \sum_{i \in [L]} \frac{1}{n_i}. \tag{10}$$

Let $\delta = \frac{\epsilon(1-\kappa^*)^2}{2}$. Since $\widehat{\kappa}$ satisfies a **(RC)**, there exists constants $A_{1,\delta}, A_{2,\delta} > 0$ such that for large enough $\boldsymbol{n}$,

$$|\widehat{\kappa} - \kappa^*| \leqslant A_{1,\delta}\gamma_{\boldsymbol{n}} + \delta$$

with probability at least $1 - A_{2,\delta} \sum_{i \in [L]} \frac{1}{n_i}$. Since $F \neq H$, $\kappa^* < 1$ by Proposition 2, so we can let $\boldsymbol{n}$ large enough so that

$$\frac{1}{(1-\kappa^*)(1-\widehat{\kappa})} \leqslant 2\frac{1}{(1-\kappa^*)^2}$$

with high probability. Then, on this same event, for large enough $\boldsymbol{n}$,

$$\begin{aligned}
\left|\frac{1}{1-\kappa^*} - \frac{1}{1-\widehat{\kappa}}\right| &\leqslant \frac{A_{1,\delta}\gamma_{\boldsymbol{n}} + \delta}{(1-\kappa^*)(1-\widehat{\kappa})} \\
&\leqslant 2\frac{A_{1,\delta}\gamma_{\boldsymbol{n}} + \delta}{(1-\kappa^*)^2} \\
&\leqslant 2\frac{A_{1,\delta}\gamma_{\boldsymbol{n}}}{(1-\kappa^*)^2} + \epsilon.
\end{aligned}$$

Thus, we obtain the claim.

We can write $G = \alpha F + (1-\alpha)H$ with $\alpha \geqslant 1$. Then, by the triangle inequality,

$$\begin{aligned}
|\widehat{G} - G| &= |\widehat{\alpha}\widehat{F} + (1-\widehat{\alpha})\widehat{H} - \alpha F - (1-\alpha)H| \\
&\leqslant |\widehat{\alpha}\widehat{F} - \alpha F| + |(1-\widehat{\alpha})\widehat{H} - (1-\alpha)H| \\
&= |\widehat{\alpha}\widehat{F} - \widehat{\alpha}F + \widehat{\alpha}F - \alpha F| + |(1-\widehat{\alpha})\widehat{H} - (1-\widehat{\alpha})H + (1-\widehat{\alpha})H - (1-\alpha)H| \\
&\leqslant |\widehat{\alpha}||\widehat{F} - F| + |\widehat{\alpha} - \alpha| + |1-\widehat{\alpha}||\widehat{H} - H| + |\widehat{\alpha} - \alpha|.
\end{aligned}$$

Since $\widehat{F}$ satisfies a **(UDI)**, $\widehat{H}$ satisfies a **(UDI)**, inequality (10) holds, and $|\widehat{\alpha}|$ and $|1-\widehat{\alpha}|$ are bounded in probability, the result follows by an application of a union bound and picking the ϵs in the uniform deviation inequalities appropriately for each term. ■

Lemma 10 gives sufficient conditions under which $\widehat{\kappa}$ satisfies **(RC)**.

Lemma 10 *Let F and H be distributions such that $F \neq H$. If*

- *F satisfies **(SC)** with respect to H,*

- $\widehat{F}$ *satisfies* ***(UDI)****, and*
- $\widehat{H}$ *satisfies* ***(UDI)****,*

then $\widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$ *satisfies* ***(RC)****.*

Proof For abbreviation, let $\kappa^* = \kappa(F \,|\, H)$ and $\widehat{\kappa} = \widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$.

We first prove the upper bound. F satisfies **(SC)** with respect to H, so there exists a distribution G such that $F = (1-\gamma)G + \gamma H$ for some $\gamma \in [0,1)$ and $\text{supp}(H) \nsubseteq \text{supp}(G)$. Therefore, we have that G is irreducible with respect to H and, by Proposition 2, $\kappa^* = \gamma$.

Let $\delta > 0$ (to be chosen later). Since by hypothesis $\widehat{F}$ and $\widehat{H}$ satisfy **(UDI)**, there exist constants $A_{1,\delta}, A_{2,\delta} > 0$ such that for large enough $\boldsymbol{n}$, with probability at least $1 - A_{1,\delta}[\sum_{i\in[L]} \frac{1}{n_i}]$, for all $E \in \mathcal{E}$,

$$|\widehat{F}(E) - F(E)| < A_{2,\delta}\gamma_{\boldsymbol{n}} + \delta \tag{11}$$

$$|\widehat{H}(E) - H(E)| < A_{2,\delta}\gamma_{\boldsymbol{n}} + \delta. \tag{12}$$

Without loss of generality, let $A_{1,\delta}, A_{2,\delta} > 1$.

Pick $R \in \mathcal{E}$ such that $H(R) > 0$. By inequality (12), there exists $\boldsymbol{N}_1$ such that $\boldsymbol{n} \geqslant \boldsymbol{N}_1$ implies that $\widehat{H}(R) - \gamma_{\boldsymbol{n}} > 0$ with high probability. This implies that for $\boldsymbol{n} \geqslant \boldsymbol{N}_1$, $\widehat{\kappa}$ is finite. Let $\epsilon > 0$. By definition of $\widehat{\kappa}$, there exists $E \in \mathcal{E}$ such that

$$\frac{\epsilon}{2} + \widehat{\kappa} \geqslant \frac{\widehat{F}(E) + \gamma_{\boldsymbol{n}}}{(\widehat{H}(E) - \gamma_{\boldsymbol{n}})_+}.$$

Since $\widehat{\kappa}$ is finite, we have that $\widehat{H}(E) > \gamma_{\boldsymbol{n}}$ and $H(E) > 0$. Then,

$$\begin{aligned}
\frac{\epsilon}{2} + \widehat{\kappa} &\geqslant \frac{\widehat{F}(E) + \gamma_{\boldsymbol{n}}}{\widehat{H}(E) - \gamma_{\boldsymbol{n}}} \\
&\geqslant \frac{F(E) - (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} - \delta}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} \\
&\geqslant \frac{\gamma H(E)}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} - \frac{(A_{2,\delta} - 1)\gamma_{\boldsymbol{n}}}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} - \frac{\delta}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} \\
&\geqslant \frac{\gamma H(E)}{H(E)} - \frac{(A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta}{H(E)} - \frac{(A_{2,\delta} - 1)\gamma_{\boldsymbol{n}}}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} - \frac{\delta}{H(E) + (A_{2,\delta} - 1)\gamma_{\boldsymbol{n}} + \delta} \\
&\geqslant \kappa^* - 2\frac{(A_{2,\delta} - 1)\gamma_{\boldsymbol{n}}}{H(E)} - 2\frac{\delta}{H(E)}
\end{aligned}$$

where in the second to last inequality we used the elementary fact that if $a, b, c > 0$ and $a \leqslant b$, then $\frac{a}{b+c} \geqslant \frac{a}{b} - \frac{c}{b}$. Picking $\delta = \frac{H(E)\epsilon}{4}$, we obtain the upper bound.

The proof of the other direction of the inequality is very similar to the proof of Theorem 2 in Scott (2015). By hypothesis, F satisfies **(SC)** with respect to H, so there exists a distribution G such that $F = (1-\gamma)G + \gamma H$ for some $\gamma \in [0,1)$ and $\text{supp}(H) \nsubseteq \text{supp}(G)$. Therefore, we have that G is irreducible with respect to H and, by Proposition 2, $\kappa^*(F \,|\, H) = \gamma$. For

abbreviation, let $\kappa^* = \kappa^*(F \,|\, H)$ and $\widehat{\kappa} = \widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$. Since $\text{supp}(H) \nsubseteq \text{supp}(G)$, there exists an open set O such that

$$\frac{F(O)}{H(O)} = (1-\gamma)\frac{G(O)}{H(O)} + \gamma = \kappa^*.$$

Then, since $\mathcal{E}$ contains a generating set for the standard topology on $\mathbb{R}^d$, there exists $E \in \mathcal{E}$ such that

$$\frac{F(E)}{H(E)} = \kappa^*.$$

Let $\delta > 0$ such that $\delta \leqslant \frac{1}{4}H(E)$. Since by hypothesis $\widehat{F}$ and $\widehat{H}$ satisfy **(UDI)**, there exist constants $A_{3,\delta}, A_{4,\delta} > 0$ such that for large enough $\boldsymbol{n}$, with probability at least $1 - A_{3,\delta}[\sum_{i\in[L]} \frac{1}{n_i}]$,

$$\begin{aligned}\widehat{\kappa} &\leqslant \frac{F(E) + A_{4,\delta}\gamma_{\boldsymbol{n}} + \delta}{(H(E) - A_{4,\delta}\gamma_{\boldsymbol{n}} - \delta)_+} \\ &\leqslant \frac{F(E) + \epsilon}{(H(E) - \epsilon)_+}\end{aligned}$$

where $\epsilon = 2A_{4,\delta}\gamma_{\boldsymbol{n}} + \delta$. The rest of the proof is identical to the proof of Theorem 2 from Scott (2015) and, therefore, we omit it. ■

The following theorem gives sufficient conditions under which a ResidueHat estimator satisfies **(UDI)**. It is the basis of Proposition 4.

Lemma 11 *If $P_1, \ldots, P_L$ satisfy **(A″)** and $\widehat{G}$ is a ResidueHat estimator of order k of a distribution $G \in \text{conv}(P_1, \ldots, P_L)$, then $\widehat{G}$ satisfies **(UDI)**.*

Proof Let $\widehat{G} \longleftarrow \text{ResidueHat}(\widehat{F} \,|\, \widehat{H})$ where $G \longleftarrow \text{ResidueHat}(F \,|\, H)$, $F \neq H$, $F, H \in \text{conv}(P_1, \ldots, P_L)$ and $\widehat{F}, \widehat{H}$ are ResidueHat estimators of F and H respectively. We use induction on k. Suppose $k = 0$. Then, $\widehat{F}$ and $\widehat{H}$ are empirical distributions. Therefore, the VC inequality applies to $\widehat{F}$ and $\widehat{H}$. Consequently, $\widehat{F}$ and $\widehat{H}$ satisfy **(UDI)**. Since $P_1, \ldots, P_L$ satisfy **(A″)**, $F, H \in \text{conv}(P_1, \ldots, P_L)$ and $F \neq H$, by Lemma 8, F satisfies **(SC)** with respect to H. Then, by Lemma 10, $\widehat{\kappa}(\widehat{F} \,|\, \widehat{H})$ satisfies **(RC)**. Then, all of the assumptions of Lemma 9 are satisfied, so $\widehat{G}$ satisfies **(UDI)**. Note that $G \in \text{conv}(P_1, \ldots, P_L)$ by Proposition 3.

The inductive step ($k > 0$) follows by similar reasoning. The difference is that instead of applying the VC inequality to $\widehat{F}$ and $\widehat{H}$, we use the fact that $\widehat{F}$ and $\widehat{H}$ are ResidueHat estimators of order $k-1$ and, therefore, satisfy **(UDI)** by the inductive hypothesis. ■

Proof [Proof of Proposition 4] Let $0 < \delta < \epsilon$. By Lemma 11, $\widehat{G}$ satisfies **(UDI)**. Consequently, there exist constants $A_{1,\delta}, A_{2,\delta} > 0$ such that for large enough $\boldsymbol{n}$ with probability at least $1 - A_{1,\delta}\sum_{i\in[L]} \frac{1}{n_i}$, $\widehat{G}$ satisfies for every $E \in \mathcal{E}$,

$$|\widehat{G}(E) - G(E)| \leqslant A_{2,\delta}\gamma_{\boldsymbol{n}} + \delta = A_{2,\delta} \sum_{i\in[L]} \epsilon_i(\frac{1}{n_i}) + \delta \longrightarrow \delta < \epsilon.$$

■

E.2. Demixing Mixed Membership Models

In this section, we prove our main estimation result for demixing mixed membership models, i.e., Theorem 5. First, in Section E.2.1, we present an important lemma for FaceTestHat. Second, in Section E.2.2, we present an empirical version of Demix and prove Theorem 5.

E.2.1. The FaceTestHat Algorithm

The following establishes that FaceTestHat behaves as desired.

Lemma 12 *Let $\epsilon \in (0,1)$. For all $j \in [K]$, let $Q_j = \boldsymbol{\eta}_j^T \boldsymbol{P}$ and $\boldsymbol{\eta}_j \in \Delta_K$ such that every $\boldsymbol{\eta}_j$ lies in the relative interior of the same face of Δ_K. Let $P_1, \ldots, P_K$ satisfy **(A″)**, and $Q_1, \ldots, Q_K \in \mathrm{conv}(P_1, \ldots, P_K)$ be distinct. Let $\widehat{Q}_i$ be a ResidueHat estimate of Q_i $\forall i \in [K]$.*

1. *With probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, if FaceTestHat($\widehat{Q}_1, \cdots, \widehat{Q}_K \,|\, \epsilon$) returns 1, then $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ are in the relative interior of the same face.*
2. *Let $\kappa^*_{i,j} = \kappa^*(Q_i \,|\, Q_j)$. If $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ are in the relative interior of the same face and $\min_{i,j} \kappa^*_{i,j} > \epsilon$, then with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, FaceTestHat($\widehat{Q}_1, \cdots, \widehat{Q}_K \,|\, \epsilon$) returns 1.*

Proof Let $\epsilon > 0$, $\kappa^*_{i,j} = \kappa^*(Q_i \,|\, Q_j)$ and $\widehat{\kappa}_{i,j} = \widehat{\kappa}(\widehat{Q}_i \,|\, \widehat{Q}_j)$. Since $P_1, \ldots, P_K$ satisfy **($\mathbf{A}''$)** and $Q_i \neq Q_j$, by Lemma 8, Q_i satisfies **(SC)** wrt Q_j. Since $\widehat{Q}_i$ and $\widehat{Q_j}$ are ResidueHat estimators, $\widehat{Q}_i$ and $\widehat{Q}_j$ satisfy **(UDI)** (Lemma 11). Then, by Lemma 10 $\widehat{\kappa}_{i,j}$ satisfies **(RC)**.

1. We prove the contrapositive. Suppose that $Q_1, \ldots, Q_K$ are not in the relative interior of the same face. Then, by Proposition 8, FaceTest($Q_1, \ldots, Q_K$) returns 0, which occurs if and only if there exist $i \neq j$ such that $\kappa^*_{i,j} = 0$. Since $\widehat{\kappa}_{i,j}$ satisfies **(RC)**, as $\boldsymbol{n} \longrightarrow \infty$, with probability tending to 1, $\widehat{\kappa}_{i,j} \longrightarrow 0$. This completes the proof.
2. If $\min_{i,j} \kappa^*_{i,j} > \epsilon$, then as $\boldsymbol{n} \longrightarrow \infty$, with probability tending to 1, $\min_{i,j} \widehat{\kappa}_{i,j} > \epsilon$.

■

E.2.2. The DemixHat Algorithm

The DemixHat algorithm (see Algorithm 11) differs from the Demix algorithm in that *(i)* it requires the specification of a constant $\epsilon \in (0,1)$ and *(ii)* it only uses the two-sample κ^* operator. In the interest of clarity, we state the population version of the algorithm DemixHat, which we call Demix2. The only difference between Demix and Demix2 is that line 7 in Demix has been replaced with lines 6-8 in Demix2.

Lemma 13 establishes that it is possible to replace line 7 of the Algorithm 4 with the sequence of applications of the two-sample κ^* in lines 6-8 of Algorithm 16, without changing the conclusion of Theorem 2.

Lemma 13 *Let $\{i_1, \ldots, i_K\} \subset [L]$ be distinct indices. Let $P_1, \ldots, P_L$ be jointly irreducible, $(Q_1, \ldots, Q_{K-1})$ be a permutation of $(P_{i_1}, \ldots, P_{i_{K-1}})$ and $Q^1_K \in \mathrm{conv}(P_{i_1}, \ldots, P_{i_K})^\circ$. Define*

Algorithm 16 Demix2$(S_1, \ldots, S_K)$

Input: $S_1, \ldots, S_K$ are distributions

if $K = 2$ **then**
  return $(\text{Residue}(S_1 \,|\, S_2), \text{Residue}(S_2 \,|\, S_1))^T$
else
  $(R_2, \ldots, R_K)^T \longleftarrow \text{FindFace}(S_1, \ldots, S_K)$
  $(Q_1, \ldots, Q_{K-1})^T \longleftarrow \text{Demix2}(R_2, \ldots, R_K)$
  for $i = 1, \ldots, K-1$ **do**
    $Q_K \longleftarrow \text{Residue}(Q_K \,|\, Q_i)$
  end for
  return $(Q_1, \ldots, Q_K)^T$
end if

the sequence

$$Q_K^i \longleftarrow \mathit{Residue}(Q_K^{i-1} \,|\, Q_{i-1});$$

then, $Q_K^K = P_{i_K}$.

Proof Relabel the distributions so that $Q_j = P_j$. Let $\boldsymbol{\mu}_i$ denote the mixture proportion of Q_K^i and $\boldsymbol{e}_j$ the mixture proportion of P_j. Write $\boldsymbol{\mu}_1 = \sum_{i=1}^K \alpha_i \boldsymbol{e}_i$. We claim that $\boldsymbol{\mu}_k = \frac{\sum_{i \geqslant k} \alpha_i \boldsymbol{e}_i}{\sum_{i \geqslant k} \alpha_i}$ for all $k \leqslant K$. We prove this inductively. The base case $k = 1$ follows since $\sum_{i \geqslant 1} \alpha_i = 1$. Next, we prove the inductive step. Suppose that $\boldsymbol{\mu}_{k-1} = \frac{\sum_{i \geqslant k-1} \alpha_i \boldsymbol{e}_i}{\sum_{i \geqslant k-1} \alpha_i}$. By Proposition 3, the mixture proportion of Q_K^k, $\boldsymbol{\mu}_k$, is the residue of $\boldsymbol{\mu}_{k-1}$ with respect to $\boldsymbol{e}_{k-1}$. By statement 1 of Lemma 2, we can write

$$\begin{aligned}
\boldsymbol{\mu}_k &= \boldsymbol{e}_{k-1} + \alpha^*(\boldsymbol{\mu}_{k-1} - \boldsymbol{e}_{k-1}) \\
&= \frac{[\sum_{i \geqslant k-1} \alpha_i (1 - \alpha^*) + \alpha^* \alpha_{k-1}] \boldsymbol{e}_{k-1} + \alpha^* \sum_{i \geqslant k} \alpha_i \boldsymbol{e}_i}{\sum_{i \geqslant k-1} \alpha_i}
\end{aligned}$$

where

$$\alpha^* = \frac{1}{1 - \kappa^*(\boldsymbol{\mu}_{k-1} \,|\, \boldsymbol{e}_{k-1})}$$

and we have used the inductive hypothesis $\boldsymbol{\mu}_{k-1} = \frac{\sum_{i \geqslant k-1} \alpha_i \boldsymbol{e}_i}{\sum_{i \geqslant k-1} \alpha_i}$. α^* is the value of the following optimization problem (in statement 1 of Lemma 2):

$$\max(\alpha \geqslant 1 \,|\, \exists G, G = \boldsymbol{\mu}_{k-1} + \alpha(\boldsymbol{e}_{k-1} - \boldsymbol{\mu}_{k-1})).$$

Inspection of the above optimization problem reveals that $\alpha^* = \frac{\sum_{i \geqslant k-1} \alpha_i}{\sum_{i \geqslant k} \alpha_i}$. Plugging this into the above equation gives $\boldsymbol{\mu}_k = \frac{\sum_{i \geqslant k} \alpha_i \boldsymbol{e}_i}{\sum_{i \geqslant k} \alpha_i}$. This establishes the claim.

Setting $k = K$, it follows that $\mu_K = \frac{\alpha_K \boldsymbol{e}_K}{\alpha_K} = \boldsymbol{e}_K$. ■

Corollary 2 *Let $P_1, \ldots, P_L$ be jointly irreducible and $\mathbf{\Pi}$ have full column rank. Then, with probability* 1*, Demix2($\tilde{\boldsymbol{P}}$) returns a permutation of $\boldsymbol{P}$.*

Proof [Proof of Theorem 5] Note that every estimator of a distribution in the DemixHat algorithm is a ResidueHat estimator since *(i)* the Demix2 algorithm 16 only considers distributions that are in $\text{conv}(P_1, \ldots, P_L)$ and *(ii)* only computes $\text{Residue}(F \,|\, H)$ if $F \neq H$. To see why *(ii)* is true, consider: the Demix2 algorithm computes $\text{Residue}(\cdot \,|\, \cdot)$ at lines 2 and 7 in Demix2 and line 3 in FindFace. In the proof of Theorem 2, we showed that $S_1, \ldots, S_K$ are always linearly independent and therefore distinct. This implies that in lines 2 and 7 in Demix2 and line 3 in FindFace, the residue function is called on distinct distributions. Thus, every estimator of a distribution of the DemixHat algorithm satisfies the assumptions of Lemma 11.

First, we argue that the order of the ResidueHat estimators is bounded; this implies that the constants in the uniform deviation inequalities associated with the ResidueHat estimators are bounded. We give a very loose bound. DemixHat calls itself at most $L-1$ times and in each call recurses on at most $L-1$ ResidueHat estimators and calculates at most $L-1$ more ResidueHat estimators. Therefore, each ResidueHat estimator has order at most $(L-1)^3$.

Second, let A_i denote the event that DemixHat recurses on i distributions lying in the relative interior of an i-face in the $(L-i)$th recursive call. We show that the event $\cap_{i=2}^{L-1} A_i$ occurs with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$. Consider A_{L-1}. Let $\widehat{R}_i^{(n)}$ denote the estimate of the ith distribution in line 3 in the nth iteration of the for loop in Algorithm 12 and let $R_i^{(n)}$ denote the corresponding distribution. Let $\kappa^*_{i,j,n} = \kappa^*(R_i^{(n)} \,|\, R_j^{(n)})$. From the proof of Theorem 2, there exists an integer $N_1 \geqslant 0$ such that for $n \geqslant N_1$, $R_i^{(n)}$ lies in the relative interior of the same face for all $i = 2, \ldots, L$. Further, using the notation from the proof of Theorem 2, we have that the mixture proportions of the $R_i^{(n)}$s, i.e., the $\boldsymbol{\mu}_i^{(n)}$s, converge to a common $\boldsymbol{\lambda}$ on this face, i.e., for all $i = 2, \ldots, L$, $\left\|\boldsymbol{\mu}_i^{(n)} - \boldsymbol{\lambda}\right\| \longrightarrow 0$. Thus, by statement 3 of Lemma 4, for all $i \neq j \in [L] \backslash \{1\}$ $\kappa^*(\boldsymbol{\mu}_i^{(n)} \,|\, \boldsymbol{\mu}_j^{(n)}) \longrightarrow 1$. Hence, there exists $N_2 \geqslant N_1$ such that $\kappa^*(\boldsymbol{\mu}_i^{(N_2)} \,|\, \boldsymbol{\mu}_j^{(N_2)}) > \epsilon$ for all $i \neq j$. By statement 1 of Lemma 12 and a union bound argument, with probability increasing to 1, $\text{FaceTestHat}(\widehat{R}_2^{(n)}, \ldots, \widehat{R}_L^{(n)} \,|\, \epsilon)$ returns 0 for all $n < N_1$ since $R_2^{(n)}, \ldots, R_L^{(n)}$ are not on the relative interior of the same face. Thus, with probability tending to 1, FaceTest does not make the mistake to return 1 before the distributions $R_2^{(n)}, \ldots, R_L^{(n)}$ are on the relative interior of the same face. By statement 2 of Lemma 12, with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, $\text{FaceTestHat}(\widehat{R}_2^{(N_2)}, \ldots, \widehat{R}_L^{(N_2)} \,|\, \epsilon)$ returns 1. Hence, with probability increasing to 1, the event A_{L-1} occurs. Applying the same argument to A_i for $i < L-1$ and taking the union bound shows that $\cap_{i=2}^{L-1} A_i$ occurs with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$.

Now, we can complete the proof. Under the assumptions of Theorem 2, there is a permutation σ such that for each distribution Q_i estimated by $\widehat{Q}_i$, $P_{\sigma(i)} = Q_i$. By Proposition 4, as $\boldsymbol{n} \longrightarrow \infty$, $\widehat{Q}_i$ converges uniformly to Q_i. The result follows. ■

Algorithm 17 VertexTestHat($\mathbf{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T, (\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$)

Form the matrix $\widehat{\boldsymbol{M}}_{i,j} \coloneqq \widehat{\kappa}(\tilde{P}_i^\dagger \,|\, \widehat{Q}_j)$
Let $|\mathbf{\Pi}^+|$ denote the number of nonzero entries in $\mathbf{\Pi}^+$
Form the matrix $\widehat{\boldsymbol{Z}}$ by making the $|\mathbf{\Pi}^+|$ largest entries of $\widehat{\boldsymbol{M}}$ equal to 1 and the rest of its entries equal to 0
Use any algorithm that finds a permutation matrix $\boldsymbol{C}$ such that $\widehat{\boldsymbol{Z}}\boldsymbol{C} = \mathbf{\Pi}^+$ (if it exists)

if such a permutation matrix $\boldsymbol{C}$ exists **then**
  return $(1, \boldsymbol{C}^T)$
else
  return $(0, \mathbf{0})$
end if

E.3. Classification with Partial Labels

In this section, we prove Theorem 6. To begin, we briefly sketch an argument that one can reduce any instance of a partial label model satisfying **(B3)** and **(A)** to an instance of a partial label model that also satisfies **(D)**. Let $J = \{i : \mathbf{\Pi}^+_{i,:} = \boldsymbol{e}_j^T \text{ for some } j \in [L]\} = \{j_1, \ldots, j_k\}$, the set of indices of contaminated distributions that are equal to some base distribution. Compute Residue($\tilde{P}_i \,|\, \tilde{P}_{j_1}$) for $i \in [L] \backslash J$ if there is l such that $\mathbf{\Pi}^+_{i,l} = \mathbf{\Pi}^+_{j_1,l} = 1$. Replace $\tilde{P}_i$ with Residue($\tilde{P}_i \,|\, \tilde{P}_{j_1}$) (and call it $\tilde{P}_i$ for simplicity of presentation). Update $\mathbf{\Pi}^+$ and remove j_1 from J. Repeat this procedure until J is empty. Then, there will be $(L - |J|)$ $\tilde{P}_i$ lying in a $(L - |J|)$-face of Δ_L that are not equal to any of the base distributions and the other contaminated distributions will be equal to base distributions. Then, it suffices to solve the instance of the partial label model on the $(L - |J|)$-face, which satisfies **(D)**.

Next, we introduce VertexTestHat (Algorithm 17), an empirical version of VertexTest, and prove that it satisfies a useful consistency property.

Lemma 14 *Suppose that $P_1, \ldots, P_L$ satisfy **(A″)**, $\mathbf{\Pi}$ has full column rank, the columns of $\mathbf{\Pi}^+$ are unique and $\mathbf{\Pi}^+$ satisfies **(D)**. Let $\widehat{Q}_1, \ldots, \widehat{Q}_L$ be ResidueHat estimators of $Q_1, \ldots, Q_L$, respectively. Suppose that $(Q_1, \ldots, Q_L)$ is a permutation of $(P_1, \ldots, P_L)$. Then, with probability tending to 1 as $\boldsymbol{n} \longrightarrow \infty$, VertexTestHat($\mathbf{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T, (\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$) returns a permutation matrix $\boldsymbol{C}$ such that $\forall i$, $\boldsymbol{C}_{i,:}(\widehat{Q}_1, \ldots, \widehat{Q}_L)^T$ is a ResidueHat estimator of P_i.*

Proof Define $\widehat{\kappa}_{i,j} \coloneqq \widehat{\kappa}(\tilde{P}_i^\dagger \,|\, \widehat{Q}_j)$ and $\kappa^*_{i,j} \coloneqq \kappa^*(\tilde{P}_i \,|Q_j)$. We claim that $\widehat{\kappa}_{i,j}$ satisfies a **(RC)**. Since $P_1, \ldots, P_L$ satisfy **(A″)** and by assumption **(D)** $Q_j \neq \tilde{P}_i$, by Lemma 8, $\tilde{P}_i$ satisfies **(SC)** wrt Q_j. Since $\tilde{P}_i^\dagger$ is an empirical distribution, $\tilde{P}_i^\dagger$ satisfies a **(UDI)**. Since $\widehat{Q}_j$ is a ResidueHat estimator, $\widehat{Q}_j$ satisfies a **(UDI)** by Lemma 11. Therefore, the hypotheses of Lemma 10 are satisfied and $\widehat{\kappa}_{i,j}$ satisfies a **(RC)**.

Form the matrix $Z_{i,j} = \mathbf{1}_{\{\kappa^*(\tilde{P}_i \,|\, Q_j) > 0\}}$ as in Algorithm 9. Since $Q_1, \ldots, Q_L$ are a permutation of $P_1, \ldots, P_L$, $\boldsymbol{Z}$ is formed by permuting the columns of $\mathbf{\Pi}^+$ appropriately. Thus, there are $|\mathbf{\Pi}^+|$ (i, j) pairs such that $\kappa_{i,j} > 0$ and the rest are such that $\kappa_{i,j} = 0$. Then, using Lemma 10 and a union bound, with probability tend-

ing to 1 as $\boldsymbol{n} \longrightarrow \infty$, $\widehat{\kappa}_{i,j}$ is among the $|S|$ largest values in the matrix $\widehat{M}$ if and only if $\kappa_{i,j} > 0$. On this event, VertexTestHat$(\boldsymbol{\Pi}^+, (\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger)^T, (\widehat{Q}_1, \ldots, \widehat{Q}_L)^T)$ and VertexTest$(\boldsymbol{\Pi}^+, (\tilde{P}_1, \ldots, \tilde{P}_M)^T, (Q_1, \ldots, Q_L)^T)$ return the same output. Since $(Q_1, \ldots, Q_L)$ is a permutation of $(P_1, \ldots, P_L)$ by hypothesis, by Lemma 7 VertexTest$(\boldsymbol{\Pi}^+, (\tilde{P}_1, \ldots, \tilde{P}_M)^T, (Q_1, \ldots, Q_L)^T)$ returns $(1, \boldsymbol{C}^T)$ such that $\boldsymbol{C}^T(Q_1, \ldots, Q_L)^T = \boldsymbol{P}$. The result follows. ■

Proof [Proof of Theorem 6] Let $(\widehat{Q}_1, \ldots, \widehat{Q}_L) \longleftarrow$ DemixHat$(\tilde{P}_1^\dagger, \ldots, \tilde{P}_M^\dagger \,|\, \epsilon)$. By Theorem 5, w.p. tending towards 1 as $\boldsymbol{n} \longrightarrow \infty$, there exists a permutation $\sigma : [L] \longrightarrow [L]$ such that for every $i \in [L]$,

$$\sup_{E \in \mathcal{E}} |\widehat{Q}_i(E) - P_{\sigma(i)}(E)| < \delta.$$

From the proof of Theorem 5, each $\widehat{Q}_i$ is a ResidueHat estimator. The assumptions of Lemma 14 are satisfied. The result follows immediately from Lemma 14. ■

Appendix F. Previous Results

Lemma 15 (Lemma A.1 (Blanchard and Scott, 2014)) *The maximum operation in the definition of κ^* and $\widehat{\kappa}$ (lines (3) and (6), respectively) is well-defined, that is, the outside supremum is attained at at least one point.*

Lemma 16 (Lemma B.1 (Blanchard and Scott, 2014)) *If Π satisfies **(B1)**, then $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L$ are linearly independent. If $P_1, \ldots, P_L$ are jointly irreducible, then they are linearly independent. If $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L$ are linearly independent and $P_1, \ldots, P_L$ are linearly independent, then $\tilde{P}_1, \ldots, \tilde{P}_L$ are linearly independent.*

Appendix G. Experiments

In this Section, we perform experiments that suggest that joint irreducibility of $P_1, \ldots, P_L$ is a reasonable assumption. In particular, our experiments suggest that on the datasets in question, **(A″)** holds (which is a strictly stronger condition than joint irreducibility). We consider three datasets: classes 1, 2, and 3 of MNIST (LeCun et al., 1998), the Iris dataset (Fisher, 1936), and the Breast Cancer Wisconsin (Diagnostic) Data Set (Dheeru and K. Taniskidou, 2017). We use the Spectral Support Estimation algorithm (De Vito et al., 2010; Rudi et al., 2014) to estimate the support of each class in each dataset. We split each dataset into training, validation, and test sets, applying the algorithm to the training set, using the validation set to pick the hyperparameters, and evaluating the performance on the test set. We average our results over 60 trials where in each trial we randomly permute the dataset, thus altering the training, validation, and test sets. Let $\widehat{S}_i$ denote an estimate of the support of class i. Tables 1, 3, and 5 display an estimate of the probability that a point sampled from P_i belongs to the estimate of the support $\widehat{S}_i$. They indicate that the Spectral Support Estimation has reasonably good performance in producing $\widehat{S}_i$s containing the support of the associated class. Tables 2, 4, and 6 use the $\widehat{S}_i$ to estimate the quantity

$\Pr_{\boldsymbol{x}\sim P_i}(\boldsymbol{x} \in \cup_{j\neq i} \mathrm{supp}(P_j))$, which must be strictly less than 1 for **(A″)** to hold. We find that our estimates are considerably less than 1, which suggests that joint irreducibility holds on these datasets.

	$i=1$	$i=2$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\widehat{S}_i)$	0.87	0.89

Table 1: Cancer Support Results.

	$i=1$	$i=2$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\cup_{j\neq i}\widehat{S}_j)$	0.18	0.38

Table 2: Cancer Separability Results.

	$i=1$	$i=2$	$i=3$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\widehat{S}_i)$	0.86	0.84	0.84

Table 3: Iris Support Results.

	$i=1$	$i=2$	$i=3$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\cup_{j\neq i}\widehat{S}_j)$	0.0	0.17	0.19

Table 4: Iris Separability Results.

	$i=1$	$i=2$	$i=3$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\widehat{S}_i)$	0.98	0.87	0.83

Table 5: MNIST Support Results.

	$i=1$	$i=2$	$i=3$
$\widehat{\Pr}_{\boldsymbol{x}\sim P_i}(\boldsymbol{x}\in\cup_{j\neq i}\widehat{S}_j)$	0.08	0.17	0.14

Table 6: MNIST Separability Results.

References

S. Arora, R. Ge, and A. Moitra. Learning topic models–going beyond svd. *Foundations of Computer Science*, 2012.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. *International Conference on Machine Learning*, 2013.

S. Axler. *Linear Algebra Done Right*. Springer, 3rd edition, 2015.

L. Berkman, B. H. Singer, and K. Manton. Black/white differences in health status and mortality among the Elderly. *Demography*, 26:661–678, 1989.

G. Blanchard and C. Scott. Decontamination of mutually contaminated models. *International Conference of Artificial Intelligence and Statistics*, 2014.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10: 2780–2824, 2016.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning research*, 3:993–1022, 2003.

J. Cid-Sueiro. Proper losses for learning from partial labels. *Advances in Neural Information Processing Systems*, 2012.

T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

R. Das, M. Zaheer, and C. Dyer. Gaussian lda for topic models with word embeddings. 2015.

E. De Vito, L. Rosasco, and A. Toigo. Spectral regularization for support estimation. *Advances in neural information processing systems*, 2010.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. Dheeru and E. K. Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

W. Ding, M. Rohban, P. Ishwar, and V. Saligrama. Topic discovery through data dependent and random projections. *International Conference on Machine Learning*, 2013.

W. Ding, M. Rohban, P. Ishwar, and V. Saligrama. Efficient distributed topic modeling with provable guarantees. *International Conference on Artificial Intelligence and Statistics*, 2014.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 2003.

R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, pages 179–188, 1936.

A. Ghosh, H. Kumar, and P.S. Sastry. Robust loss functions under label noise for deep neural networks. *AAAI*, 2017.

K. Huang, X. Fu, and N. D. Sidiropoulos. Anchor-free correlated topic modeling. *Advances in Neural Information Processing Systems*, 2016.

S. Jain, M. White, M. W. Trosset, and P. Radivojac. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016.

R. Jin and Z. Ghahramani. Learning with multiple labels. *Advances in Neural Information Processing Systems*, 2002.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *IEEE*, pages 2278–2324, 1998.

C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic modeling for short texts with auxiliary word embeddings. *ACM SIGIR conference on Research and Development in Information Retrieval*, 2016a.

F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. *Computer Vision and Pattern Recognition*, 2005.

S. Li, T. Chua, and C. Miao. Generative topic embedding: a continuous representation of documents. *ACL*, 2016b.

L.-P. Liu and T. G. Dietterich. A conditional multinomial mixture model for superset label learning. *Advances in Neural Information Processing Systems*, 2012.

L.-P. Liu and T. G. Dietterich. Learnability of the superset label learning problem. *International Conference on Machine Learning*, 32, 2014.

M. Luong, R. Socher, and C. Manning. Better word representations with recursive neural networks for morphology. *Conference on Computational Natural Language Learning*, 2013.

A. Menon, B. van Rooyen, C. Ong, and B. Williamson. Learning from corrupted binary labels via class-probability estimation. *International Conference on Machine Learning*, 2015a.

A. Menon, B. van Rooyen, C. S. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. 2015b.

A. Menon, B. van Rooyen, and N. Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv preprint*, 2016.

E. Metodiev and J. Thaler. On the topic of jets. *arXiv preprint arXiv:1802.00008*, 2018a.

E. Metodiev and J. Thaler. Jet topics: Disentangling quarks and gluons at colliders. *Physical Review Letters*, 120(24):241602, 2018b.

N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. *Advances in Neural Information Processing Systems*, 2013.

N. Nyugen and R. Caruana. Classification with partial labels. *International Conference on Knowledge Discovery and Data Mining*, 2008.

G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: a loss correction approach. *CVPR*, 2017.

J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

H. Ramaswamy, C Scott, and A Tewari. Mixture proportion estimation via kernel embeddings of distributions. *International Conference on Machine Learning*, 2016.

B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring non-negative matrices with linear programs. *Advances in Neural Information Processing Systems*, 2012.

A. Rudi, F. Odone, and E. De Vito. Geometrical and computational aspects of spectral support estimation for novelty detection. *Pattern Recognition Letters*, 2014.

T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. *International Conference on Artificial Intelligence and Statistics*, 2014.

C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. *International Conference on Artificial Intelligence and Statistics*, 2015.

C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. *Conference on Learning Theory*, 2013.

B. van Rooyen and R. Williamson. Learning in the presence of corruption. *arXiv preprint*, 2015.

B. van Rooyen, A. Menon, and R. Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in Neural Information Processing Systems*, 2015.

G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang. A correlated topic model using word embeddings. 2017.

H. Zhao, L. Du, W. Buntine, and M. Zhou. Inter and intra topic structure learning with word embeddings. 2018.