# Crowdsourcing via Annotator Co-occurrence Imputation and Provable Symmetric Nonnegative Matrix Factorization

**Shahana Ibrahim** [1] **Xiao Fu** [1]

## Abstract

Unsupervised learning of the Dawid-Skene (D&S) model from noisy, incomplete and crowdsourced annotations has been a long-standing challenge, and is a critical step towards reliably labeling massive data. A recent work takes a coupled nonnegative matrix factorization (CNMF) perspective, and shows appealing features: It ensures the identifiability of the D&S model and enjoys low sample complexity, as only the estimates of the co-occurrences of annotator labels are involved. However, the identifiability holds only when certain somewhat restrictive conditions are met in the context of crowdsourcing. Optimizing the CNMF criterion is also costly—and convergence assurances are elusive. This work recasts the pairwise co-occurrence based D&S model learning problem as a symmetric NMF (SymNMF) problem—which offers enhanced identifiability relative to CNMF. In practice, the SymNMF model is often (largely) incomplete, due to the lack of co-labeled items by some annotators. Two lightweight algorithms are proposed for co-occurrence imputation. Then, a low-complexity shifted *rectified linear unit* (ReLU)-empowered SymNMF algorithm is proposed to identify the D&S model. Various performance characterizations (e.g., missing co-occurrence recoverability, stability, and convergence) and evaluations are also presented.

## 1. Introduction

Modern machine learning systems, in particular, deep learning systems, are empowered by massive high-quality *labeled* data (Goodfellow et al., 2016; Najafabadi et al., 2015). However, massive data labeling is an arduous task—reliable data annotation requires substantial human efforts with considerable expertise, which are costly. *Crowdsourcing* techniques deal with various aspects of data labeling, ranging from crowd (annotators)-based reliable annotation acquisition to effective integration of the acquired labels (Kittur et al., 2008). Many online platforms—such as *Amazon Mechanical Turk* (AMT) (Buhrmester et al., 2011), *CrowdFlower* (Wazny, 2017), and *Clickworker* (Vakharia & Lease, 2013)—have been launched for these purposes. In platforms such as AMT, the (oftentimes self-registered) annotators do not necessarily provide reliable labels. Hence, simple integration strategies such as majority voting may work poorly (Karger et al., 2011a).

Annotation integration is a long-existing research topic in machine learning; see, e.g., (Ibrahim et al., 2019; Karger et al., 2011a; Karger et al., 2011b; Karger et al., 2013; 2014; Liu et al., 2012; Ma et al., 2018; Snow et al., 2008; Traganitis et al., 2018; Welinder et al., 2010; Zhang et al., 2016). As an unsupervised learning task, it is often tackled from a statistical generative model identification viewpoint. The *Dawid-Skene* (D&S) model (Dawid & Skene, 1979) has been widely adopted in the literature. The D&S model assumes a ground-truth label prior and assigns a "confusion" matrix to each annotator. The entries of an annotator's confusion matrix correspond to the probabilities of the correct and incorrect annotations conditioned on the ground-truth labels. Hence, annotation integration boils down to learning the model parameters of the D&S model.

Perhaps a bit surprisingly, despite its popularity, the *identifiability* of the D&S model had not been satisfactorily addressed until recent years. The model identifiability of D&S was first shown under some special cases (e.g., binary labeling cases) (Dalvi et al., 2013; Ghosh et al., 2011; Karger et al., 2013). The more general multi-class cases were discussed in (Traganitis et al., 2018; Zhang et al., 2016), assuming the availability of third-order statistics of the crowdsourced annotations. A challenge is that the third-order statistics may be difficult to estimate reliably, especially in the sample-starved regime. The work of (Ibrahim et al., 2019) used pairwise co-occurrences of the annotators' responses (i.e., second-order statistics) to identify the D&S model, which substantially improved the sample complexity,

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA. Correspondence to: Shahana Ibrahim <ibrahish@oregonstate.edu>, Xiao Fu <xiao.fu@oregonstate.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

compared to the third-order statistics-based approaches.

Using second-order statistics is conceptually appealing, yet the work in (Ibrahim et al., 2019) still faces serious challenges in handling real large-scale crowdsourcing problems.

1. **Identifiability Challenge.** The identifiability of the methods in (Ibrahim et al., 2019) hinges on a number of restrictive and somewhat unnatural assumptions, e.g., the existence of two *disjoint* groups of annotators that both contain "class specialists" for all classes.
2. **Computational Challenges.** The main algorithm in (Ibrahim et al., 2019) is based on a *coupled nonnegative matrix factorization* (CNMF) approach, which has serious scalability issues. In addition, its noise robustness and convergence properties are unclear.

### 1.1. Contributions

To overcome the challenges, we take a deeper look at the pairwise co-occurrence (second-order statistics) based D&S model identification problem and offer an alternative approach. Our contributions are as follows:

**Enhanced Identifiability.** We reformulate the pairwise annotator co-occurrence based D&S model identification problem as a *symmetric nonnegative matrix factorization* (SymNMF) problem in the presence of missing "blocks"—which are caused by the absence of some annotator co-occurrences (since not all annotators label all items). We show that if the missing co-occurrences can be correctly imputed, solving the subsequent SymNMF problem uniquely identifies the D&S model under much relaxed conditions relative to those in (Ibrahim et al., 2019).

**Co-occurrence Imputation Algorithms.** We offer two custom and recoverability-guaranteed co-occurrence imputation algorithms. First, we take advantage of the fact that annotator dispatch is under control in some crowdsourcing problems and devise a co-occurrence imputation algorithm using simple operations like singular value decomposition (SVD) and least squares (LS). Second, we consider a more challenging scenario where annotator dispatch is out of reach and some observed co-occurrences are unreliably estimated. Under this scenario, we propose an imputation criterion that is provably robust to outlying co-occurrence observations. We also propose a lightweight iterative algorithm under this setting.

**Fast and Provable SymNMF Algorithm.** To identify the D&S model from the co-occurrence-imputed SymNMF model, we propose an algorithm that is a modified version of the subspace-based SymNMF algorithm in (Huang et al., 2014). The algorithm in (Huang et al., 2014) is known for its simple updates and empirically fast convergence, but understanding to its convergence properties has been elusive. We replace the nonnegativity projection step in the algorithm by a shifted *rectified linear unit* (ReLU) operator. Consequently, we show that the new algorithm converges *linearly* to the desired D&S model parameters under some conditions—while maintaining almost the same lightweight updates. We also show that the new algorithm is provably robust to noise. Note that the SymNMF is an NP-hard problem, and analyzing the model estimation accuracy is challenging. Our convergence result fills this gap.

**Notation.** A summary of notations used in this work can be found in the supplementary material.

## 2. Background

We focus on the D&S model identification problem in the context of crowdsourced data annotation. Consider $N$ data items that are denoted as $\{\boldsymbol{f}_n\}_{n=1}^N$, where $\boldsymbol{f}_n \in \mathbb{R}^D$ is a feature vector representing the data item. The corresponding (unknown) ground-truth labels are $\{y_n\}_{n=1}^N$, where $y_n \in \{1, 2, \ldots, K\}$ and $K$ is the number of classes. These unlabeled data items are crowdsourced to $M$ annotators. Each annotator labels a subset of the $N$ items, and the subsets could be overlapped. Annotator $m$'s response to item $n$ is denoted as $X_m(\boldsymbol{f}_n) \in \{1, \ldots, K\}$. Our interest lies in integrating $\{X_m(\boldsymbol{f}_n)\}_{m\in\mathcal{I}_n}$, where $\mathcal{I}_n$ is the index set of the annotators who co-labeled item $n$, to estimate the ground-truth $y_n$ for all $n \in [N]$. Note that naïve integration methods such as majority voting often work poorly (Karger et al., 2011a; Salk et al., 2017), as the annotators are not equally reliable and the annotations from an annotator are normally (heavily) incomplete.

### 2.1. Dawid-Skene Model

Under the D&S model, the ground-truth data label and the $M$ annotators' responses are assumed to be discrete random variables (RVs), which are denoted by $Y$ and $\{X_m\}_{m=1}^M$, respectively. A key assumption is that the $X_m$'s are conditionally independent given $Y$, i.e.,

$$\mathsf{Pr}(k_1, \ldots, k_M) = \sum_{k=1}^{K} \prod_{m=1}^{M} \mathsf{Pr}(k_m|k)\mathsf{Pr}(k), \qquad (1)$$

where $k_m, k \in [K]$, and we have used the shorthand notation $\mathsf{Pr}(k_1, \ldots, k_M) = \mathsf{Pr}(X_1 = k_1, \ldots, X_M = k_M)$, $\mathsf{Pr}(k) = \mathsf{Pr}(Y = k)$ and $\mathsf{Pr}(k_m|k) = \mathsf{Pr}(X_m = k_m|Y = k)$. On the right-hand side, $\mathsf{Pr}(X_m = k_m|Y = k)$ when $k_m \neq k$ is referred to as the *confusion probability* of annotator $m$, and $\mathsf{Pr}(Y = k)$ for $k \in [K]$ is the prior probability mass function (PMF) of the ground-truth label. Identifying the D&S model, i.e., the confusion probabilities and

the prior, allows us to build up a maximum *a posteriori* probability (MAP) estimator for $y_n$.

### 2.2. Related Work - From EM to Tensor Decomposition

The work in (Dawid & Skene, 1979) offered an expectation maximization (EM) algorithm for identifying the D&S model, while no convergence or model identifiability properties were understood at the time. Later on, a number of works considered special cases of the D&S model and offered identifiability supports. For example, under the "one coin" model, the work in (Ghosh et al., 2011) established the identifiability of the D&S model via SVD. This work considered cases with binary labels and no missing annotations (i.e., all annotators label all data items). The work in (Dalvi et al., 2013) extended the ideas to more realistic settings where missing annotations exist. Around the same time, other approaches, e.g., random graph theory (Karger et al., 2013) and iteratively reweighted majority voting (Li, 2015; Li & Yu, 2014), were also used for D&S model identification. In (Welinder et al., 2010; Whitehill et al., 2009; Zhou et al., 2012; 2015), the D&S model was extended by modeling aspects such as "item difficulty" and "annotator ability". However, the identifiability of these more complex models are unclear.

The work in (Traganitis et al., 2018; Zhang et al., 2016) addressed D&S model identification with multi-class labels using third-order statistics of the annotations. The D&S model identification problem was recast as tensor decomposition problems. Consequently, the uniqueness of tensor decomposition was leveraged for provably identifying the D&S model. The key challenge lies in the sample complexity for accurately estimating the third-order statistics. The difficulty of accurately estimating the third-order statistics may make the tensor methods struggle, especially in the annotation-starved cases. Tensor decomposition may also be costly in terms of computation; see (Fu et al., 2020a;b).

### 2.3. Recent Development - Coupled NMF

Our work is motivated by a recent development in (Ibrahim et al., 2019). The work in (Ibrahim et al., 2019) used only the estimates of $\mathsf{Pr}(X_m = k_m, X_j = k_j)$'s, which are much easier to estimate compared to third-order statistics in terms of sample complexity (Han et al., 2015). Define the *confusion matrix* of annotator $m$ (denoted by $\boldsymbol{A}_m \in \mathbb{R}^{K\times K}$) and the prior PMF $\boldsymbol{\lambda} \in \mathbb{R}^K$ as follows: $\boldsymbol{A}_m(k_m, k) = \mathsf{Pr}(X_m = k_m | Y = k)$ and $\boldsymbol{\lambda}(k) = \mathsf{Pr}(Y = k)$. Then, by the conditional independence in (1), the co-occurrence matrix of annotators $m, j$ can be expressed as

$$\boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top, \tag{2}$$

where $\boldsymbol{R}_{m,j}(k_m, k_j) = \mathsf{Pr}(X_m = k_m, X_j = k_j) = \sum_{k=1}^K \boldsymbol{\lambda}(k)\boldsymbol{A}_m(k_m, k)\boldsymbol{A}_j(k_j, k)$ and $\boldsymbol{D} = \mathsf{Diag}(\boldsymbol{\lambda})$. In practice, if two annotators $m$ and $j$ co-label a number of items, then the corresponding $\boldsymbol{R}_{m,j}$ can be estimated via sample averaging, i.e.,

$$\widehat{\boldsymbol{R}}_{m,j}(k_m, k_j) = \frac{1}{|\mathcal{S}_{m,j}|} \sum_{n\in\mathcal{S}_{m,j}} \mathbb{I}\left[X_m(\boldsymbol{f}_n) = k_m, X_j(\boldsymbol{f}_n) = k_j\right], \tag{3}$$

where $\mathbb{I}[\cdot]$ is an indicator function, $k_m, k_j \in [K]$, $\mathcal{S}_{m,j} \subseteq [N]$ holds the indices of $\boldsymbol{f}_n$'s that are co-labeled by annotators $m$ and $j$, and $|\mathcal{S}_{m,j}|$ is the number of items annotators $m$ and $j$ co-labeled.

Note that not all $\boldsymbol{R}_{m,j}$'s are available since some annotators $m, j$ may not have co-labeled any items. Hence, the problem boils down to estimating $\boldsymbol{A}_m$'s and $\boldsymbol{\lambda}$ from $\boldsymbol{R}_{m,j}$'s where $(m, j) \in \boldsymbol{\Omega}$ with $m \neq j$, where $\boldsymbol{\Omega}$ is the index set of the observed pairwise co-occurrences.

The work in (Ibrahim et al., 2019) considered the following CNMF criterion:

$$\text{find } \{\boldsymbol{A}_m\}_{m=1}^M, \boldsymbol{\lambda} \tag{4a}$$

$$\text{s.t. } \boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top, \ (m, j) \in \boldsymbol{\Omega}, \tag{4b}$$

$$\boldsymbol{A}_m \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{A}_m = \boldsymbol{1}^\top, \ \boldsymbol{1}^\top \boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \boldsymbol{0}, \tag{4c}$$

where the constraints are added per the PMF interpretations of the columns of $\boldsymbol{A}_m$ and $\boldsymbol{\lambda}$. The word "coupled" comes from the fact that the co-occurrences are modeled by $\boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top$ with shared (coupled) $\boldsymbol{A}_m$'s and $\boldsymbol{A}_j$'s. It was shown in (Ibrahim et al., 2019) that under some conditions, $\boldsymbol{A}_m^\star = \boldsymbol{A}_m \boldsymbol{\Pi}$ and $\boldsymbol{D}^\star = \boldsymbol{D}\boldsymbol{\Pi}$, where $\boldsymbol{A}_m^\star$ and $\boldsymbol{D}^\star$ are from *any* optimal solution of (4) and $\boldsymbol{\Pi}$ is permutation matrix. Specifically, assume that there exist two subsets of the annotators, indexed by $\mathcal{P}_1$ and $\mathcal{P}_2$, where $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$ and $\mathcal{P}_1 \cup \mathcal{P}_2 \subseteq [M]$. Let

$$\begin{aligned} \boldsymbol{H}^{(1)} &:= [\boldsymbol{A}_{m_1}^\top, \ldots, \boldsymbol{A}_{m_{|\mathcal{P}_1|}}^\top]^\top, \\ \boldsymbol{H}^{(2)} &:= [\boldsymbol{A}_{j_1}^\top, \ldots, \boldsymbol{A}_{j_{|\mathcal{P}_2|}}^\top]^\top, \end{aligned} \tag{5}$$

where $m_t \in \mathcal{P}_1$ and $j_\ell \in \mathcal{P}_2$. The most important condition used in (Ibrahim et al., 2019) is that both $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ satisfy the *sufficiently scattered condition* (SSC) (cf. Definition 1).

**Identifiability Challenge.** One of our major motivations is that the conditions for D&S identification in (Ibrahim et al., 2019) are somewhat restrictive. To understand this, it is critical to understand the *sufficiently scattered condition* (SSC) that is imposed on $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$. SSC is widely used in the NMF literature (Fu et al., 2015; 2016; 2018; 2019; Gillis, 2020; Huang et al., 2014) and is defined as follows:

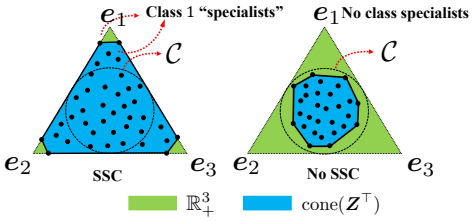


*Figure 1.* Illustration of $\boldsymbol{Z}$ satisfying the SSC and violating the SSC, respectively. The dots are the rows of $\boldsymbol{Z}$; the circle is the second-order cone $\mathcal{C}$; and the blue region is the conic hull of $\boldsymbol{Z}^\top$. To make $\boldsymbol{Z}$ satisfy the SSC, the blue region should cover the circle.

**Definition 1** *(SSC) Any nonnegative matrix $\boldsymbol{Z} \in \mathbb{R}_+^{I\times K}$ satisfies the SSC if the conic hull of $\boldsymbol{Z}^\top$ (i.e., $\mathsf{cone}(\boldsymbol{Z}^\top)$) satisfies (i) $\mathcal{C} \subseteq \mathrm{cone}\{\boldsymbol{Z}^\top\}$ where $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^K \mid \boldsymbol{x}^\top\mathbf{1} \geq \sqrt{K-1}\|\boldsymbol{x}\|_2\}$ and (ii) $\mathsf{cone}\{\boldsymbol{Z}^\top\} \not\subseteq \mathsf{cone}\{\boldsymbol{Q}\}$ for any orthonormal $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ except for the permutation matrices.*

The SSC reflects how spread are the rows of $\boldsymbol{Z}$ in the nonnegative orthant. The illustration of the SSC is shown in Fig. 1. To satisfy the SSC, some rows of $\boldsymbol{H}^{(i)}$ need to be not too far away from the extreme rays of nonnegative orthant (i.e., the unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$). This means that some rows of certain $\boldsymbol{A}_m$'s are close to be unit vectors. If $\|\boldsymbol{A}_m(k,:) - \boldsymbol{e}_k^\top\|_2$ is small, it means that $|\boldsymbol{A}_m(k,k) - 1| = |\mathsf{Pr}(X_m = k|Y = k) - 1|$ is small; i.e., annotator $m$ rarely confuses data from other classes with the ones from class $k$ and is a "class specialist" for class $k$. In other words, both $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ satisfying the SSC means that the disjoint $\mathcal{P}_1$ and $\mathcal{P}_2$ both contain "class specialists" for all $K$ classes—which may not be a trivial condition to fulfil in practice.

**Computational Challenges.** The work in (Ibrahim et al., 2019) recast the problem in (4) as a *Kullback-Leiber* (KL) divergence based model fitting problem with constraints. The iterative algorithm there often produces accurate integrated labels, but some major challenges exist. First, the method is hardly scalable. When the number of annotators grows, the runtime of the CNMF algorithm increases significantly. Second, due to the nonconvexity, it is unclear if the algorithm converges to the optimal ground-truth $\boldsymbol{A}_m$ and $\boldsymbol{D}$, even if there is no noise. Third, when there is noise, it is unclear how it affects the model identifiability, since the main theorem of (Ibrahim et al., 2019) for CNMF was derived under the ideal case where no noise is present. The work in (Ibrahim et al., 2019) offered a fast greedy algorithm for noisy cases. However, the conditions for that algorithm to work is much more restrictive, and the greedy algorithm's outputs are less accurate, as will be seen in the experiments.

## 3. Proposed Approach

Because of the appeal of its sample complexity, we offer an alternative way of using pairwise co-occurrences, while circumventing the challenges in the CNMF approach. Assume that all $\boldsymbol{R}_{m,j} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_j^\top$ are available (including the cases where $m = j$). Then, one can construct

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{R}_{1,1} & \cdots & \boldsymbol{R}_{1,M} \\ \vdots & \ddots & \vdots \\ \boldsymbol{R}_{M,1} & \cdots & \boldsymbol{R}_{M,M} \end{bmatrix} = \boldsymbol{H}\boldsymbol{H}^\top, \qquad (6)$$

where $\boldsymbol{H} = [\boldsymbol{A}_1^\top, \ldots \boldsymbol{A}_M^\top]^\top\boldsymbol{D}^{1/2}$. Note that the above is a symmetric NMF model since $\boldsymbol{H} \geq \mathbf{0}$ by the physical meaning of the $\boldsymbol{A}_m$'s and $\boldsymbol{D}$. It is known that the model is unique if $\boldsymbol{H}$ satisfies the SSC (Huang et al., 2014). Hence, we have the following:

**Proposition 1** *Assume that $\boldsymbol{H}$ in (6) satisfies the SSC, $\mathrm{rank}(\boldsymbol{H}) = K$, and that $\boldsymbol{X}$ in (6) is available. Then, all the confusion matrices and the data prior in the D&S model can be identified uniquely by SymNMF of $\boldsymbol{X}$ up to common column permutations; i.e., $\boldsymbol{A}_m^\star = \boldsymbol{A}_m\boldsymbol{\Pi}$, $\forall m \in [M]$, $\boldsymbol{\lambda}^\star = \boldsymbol{\Pi}^\top\boldsymbol{\lambda}$, where $\boldsymbol{\Pi}$ is a permutation matrix and $\boldsymbol{A}_m^\star$ denotes the $m$th column-normalized (w.r.t. the $\ell_1$ norm) block in $\boldsymbol{H}^\star$ that is any solution satisfying $\boldsymbol{X} = \boldsymbol{H}^\star(\boldsymbol{H}^\star)^\top$ with $\boldsymbol{H}^\star \geq \mathbf{0}$.*

The proof is a straightforward application of Theorem 4 in (Huang et al., 2014).

**Improved Identifiability Conditions.** Unlike in the CNMF approach, in Proposition 1, the SSC condition is imposed on $\boldsymbol{H} \in \mathbb{R}^{MK\times K}$ instead of $\boldsymbol{H}^{(i)} \in \mathbb{R}^{|\mathcal{P}_i|K\times K}$ for $i = 1, 2$. Consequently, one only needs *one* set of class specialists from all the annotators instead of *two* sets of specialists from disjoint groups of the annotators. In addition, since $\boldsymbol{H}$ is potentially much "taller" than $\boldsymbol{H}^{(i)}$ (since it is often the case that $|\mathcal{P}_i| \ll M$), the probability that it attains the SSC condition is also much higher than that of the $\boldsymbol{H}^{(i)}$'s. In fact, it was shown that, under a certain probabilistic generative model, for a nonnegative matrix $\boldsymbol{Z} \in \mathbb{R}^{I\times K}$ to satisfy the SSC with $\varepsilon$-sized error (see the detailed definition in (Ibrahim et al., 2019)) with probability of at least $1 - \mu$, one needs that $I \geq \Omega\left(\frac{(K-1)^2}{\kappa^{2(K-2)}\varepsilon^2}\log\left(\frac{K(K-1)}{\mu}\right)\right)$, where $\kappa > \varepsilon$ is a constant—which also asserts that $\boldsymbol{H}$ has a better chance to attain the SSC compared to the $\boldsymbol{H}^{(i)}$'s.

**Missing Co-occurrences.** The rationale for enhancing the D&S model identifiability using the SymNMF model in (6) is clear—but the challenges are also obvious. In particular, many blocks ($\boldsymbol{R}_{m,j}$'s) in $\boldsymbol{X}$ can be missing for different reasons. First $\boldsymbol{R}_{m,m} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_m^\top$ for $m = 1, \ldots, M$ do not have physical meaning and thus cannot be observed or

directly estimated from the data through sample averaging. Second, if annotators $m, j$ never co-labeled any items, the corresponding co-occurrence matrix $\boldsymbol{R}_{m,j}$ is missing.

Note that when $MK \gg K$, $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{H}^\top$ is a low-rank factorization model. Imputing the unobserved $\boldsymbol{R}_{m,j}$'s amounts to a *low-rank matrix completion* (LRMC) problem (Candés et al., 2011). Nonetheless, existing LRMC recoverability theory and algorithms are mostly designed under the premise that the entries (other than blocks) are missing uniformly at random—which do not cover our block missing case. In the next two subsections, we offer two co-occurrence imputation algorithms that are tailored for the special missing pattern in the context of crowdsourcing.

### 3.1. Designated Annotators-based Imputation

In crowdsourcing, annotators can sometimes be dispatched by the label requester. Hence, some annotators may be *designated* to co-label items with other annotators. To explain, consider the case where $\boldsymbol{R}_{m,n} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_n^\top$ is missing, i.e., $(m,n) \notin \boldsymbol{\Omega}$. Assume that two annotators (indexed by $\ell$ and $r$) can be designated to label items that were labeled by annotators $m$ and $n$. This way, $\boldsymbol{R}_{m,r}$, $\boldsymbol{R}_{n,\ell}$ and $\boldsymbol{R}_{\ell,r}$ can be *made* available (if there is no estimation error). Construct $\boldsymbol{C} = [\boldsymbol{R}_{m,r}^\top, \boldsymbol{R}_{\ell,r}^\top]^\top$. Consider the thin SVD of $\boldsymbol{C}$, i.e.,

$$\boldsymbol{C} = [\boldsymbol{U}_m^\top, \boldsymbol{U}_\ell^\top]^\top \boldsymbol{\Sigma}_{m,\ell,r} \boldsymbol{V}_r^\top. \tag{7}$$

When $\mathrm{rank}(\boldsymbol{A}_m) = \mathrm{rank}(\boldsymbol{D}) = K$ for all $m \in [M]$, it is readily seen that $\boldsymbol{U}_m = \boldsymbol{A}_m\boldsymbol{D}^{1/2}\boldsymbol{\Theta}$ and $\boldsymbol{U}_\ell = \boldsymbol{A}_\ell\boldsymbol{D}^{1/2}\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$ is nonsingular. Hence, one can estimate $\boldsymbol{R}_{m,n}$ via

$$\boldsymbol{R}_{m,n} = \boldsymbol{U}_m\boldsymbol{U}_\ell^{-1}\boldsymbol{R}_{n,\ell}^\top. \tag{8}$$

This simple procedure also allows us to characterize the estimation error of $\boldsymbol{R}_{m,n}$ when only a finite number of co-labeled items are available:

**Theorem 1** *Assume that $\widehat{\boldsymbol{R}}_{m,n}$ is estimated by (7)-(8) using the sample-estimated $\widehat{\boldsymbol{R}}_{m,r}$, $\widehat{\boldsymbol{R}}_{n,\ell}$ and $\widehat{\boldsymbol{R}}_{\ell,r}$ [using (3) with at least $S$ items]. Also assume that $\kappa(\boldsymbol{A}_m) \le \gamma$ and $\mathrm{rank}(\boldsymbol{A}_m) = \mathrm{rank}(\boldsymbol{D}) = K$ for all $m \in [M]$. Let $\varrho = \min_{(m,j)\in\boldsymbol{\Omega}} \sigma_{\min}(\boldsymbol{R}_{m,j})$. Suppose that $S = \Omega\left(\frac{K^2\gamma^2\log(1/\delta)}{\varrho^4}\right)$ for $\delta > 0$. Then, for any $(m,n) \notin \boldsymbol{\Omega}$, with probability of at least $1-\delta$, we have:*

$$\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} = O\left(\frac{K^2\gamma^3\sqrt{\log(1/\delta)}}{\varrho^2\sqrt{S}}\right),$$

*where $\boldsymbol{R}_{m,n} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_n^\top$ is the missing ground-truth.*

The proof can be found in the supplementary material in Sec. D. Note that the designated annotator approach can also estimate the diagonal blocks in $\boldsymbol{X}$, i.e., $\boldsymbol{R}_{m,m} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_m^\top$, by asking annotators $\ell, r$ to estimate $\boldsymbol{R}_{m,\ell}$, $\boldsymbol{R}_{m,r}$, and $\boldsymbol{R}_{\ell,r}$. The diagonal blocks can never be observed, even if every pairwise annotator co-occurrence is observed, since $\boldsymbol{R}_{m,m}$ does not have physical meaning. Hence, being able to impute the diagonal blocks is particularly important for completing the matrix $\boldsymbol{X}$.

**Remark 1** *If $\boldsymbol{R}_{m,r}$, $\boldsymbol{R}_{n,\ell}$ and $\boldsymbol{R}_{\ell,r}$ are observed, then $\boldsymbol{R}_{m,n}$ can be imputed using (7)-(8) no matter if designated annotators exist. As will be seen, this method works reasonably well even in the absence of designated annotators, especially when the number of missing co-occurrences is not large. Nonetheless, having designated annotators guarantees that every missing co-occurrence is estimated.*

### 3.2. Robust Co-occurrence Imputation

In some cases, designated annotators may not exist. More critically, the estimated co-occurrences may not be equally reliable—since the estimation accuracy of $\widehat{\boldsymbol{R}}_{m,j}$ depends on the number of items that annotators $m$ and $j$ have co-labeled [cf. Eq. (3)], which may be quite unbalanced across different co-occurrences. Under such circumstances, we propose a robust co-occurrence imputation criterion, i.e.,

$$\underset{\boldsymbol{U}_m, \boldsymbol{U}_j,\, \forall (m,j)\in\boldsymbol{\Omega}}{\text{minimize}} \sum_{(m,j)\in\boldsymbol{\Omega}} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m\boldsymbol{U}_j^\top\|_{\mathrm{F}} \tag{9a}$$

$$\text{subject to } \|\boldsymbol{U}_m\|_{\mathrm{F}} \le D,\ \|\boldsymbol{U}_j\|_{\mathrm{F}} \le D,\ \forall m, \tag{9b}$$

where $D$ is an upper bound of $\|\boldsymbol{U}_m\|_F$—which is easy to acquire in our case, as $\boldsymbol{U}_m \in \mathcal{R}(\boldsymbol{A}_m\boldsymbol{D}^{1/2})$ and $\boldsymbol{A}_m$'s and $\boldsymbol{D}$ are bounded. Our formulation can be understood as a block $\ell_2/\ell_1$-mixed norm based criterion, which is often used in robust estimation for "downweighting" outlying data; see e.g., (Fu et al., 2016; Nie et al., 2014; Xu et al., 2012).

**Stability Under Finite Sample.** Our formulation is reminiscent of matrix factorization based LRMC (see, e.g., (Sun & Luo, 2016)), but with a special block missing pattern and a co-occurrence level robustification. The existing literature of LRMC and its recoverability analysis do not cover our case. Nonetheless, we show that the proposed criterion in (9) is a sound criterion for co-occurrence imputation:

**Theorem 2** *Assume that the $\widehat{\boldsymbol{R}}_{m,j}$'s are estimated using (3) with $S_{m,j} = |\mathcal{S}_{m,j}|$ for all $(m,j) \in \boldsymbol{\Omega}$. Also assume that each $\widehat{\boldsymbol{R}}_{m,j}$ is observed with the same probability. Let $\{\boldsymbol{U}_m^*, \boldsymbol{U}_j^*\}_{(m,j)\in\boldsymbol{\Omega}}$ be any optimal solution of (9). Define*

$L = M(M-1)/2$. *Then we have*

$$\frac{1}{L}\sum_{m<j}\|\boldsymbol{U}_m^*(\boldsymbol{U}_j^*)^\top - \boldsymbol{R}_{m,j}\|_{\rm F} \le C\sqrt{\frac{MK^2\log(M)}{|\boldsymbol{\Omega}|}} + \left(\frac{1}{|\boldsymbol{\Omega}|}+\frac{1}{L}\right)\sum_{(m,j)\in\boldsymbol{\Omega}}\frac{1+\sqrt{M}}{\sqrt{S_{m,j}}}, \tag{10}$$

*with probability of at least* $1-3\exp(-M)$, *where* $C>0$.

The proof can be found in the supplementary material in Sec. E. Naturally, the criterion favors more annotators and more observed pairwise co-occurrences. A remark is that the second term on the right hand side of (10) is proportional to $\sum\|\boldsymbol{N}_{m,j}\|_{\rm F}$ where $\boldsymbol{N}_{m,j} = \widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}_{m,j}$. Unlike $\sum\|\boldsymbol{N}_{m,j}\|_{\rm F}^2$, this term is not dominated by large $\|\boldsymbol{N}_{m,j}\|_{\rm F}$'s—which reflects the criterion's robustness to badly estimated $\widehat{\boldsymbol{R}}_{m,j}$'s. Also note that the result in Theorem 2 does not include the diagonal blocks $\boldsymbol{R}_{m,m}$'s. Nonetheless, the $\boldsymbol{R}_{m,m}$'s can be easily estimated using (7)-(8) if every other $\boldsymbol{R}_{m,j}$ is (approximately) recovered.

**Iteratively Reweighted Algorithm.** We propose an iteratively reweighted alternating optimization algorithm to tackle (9). In each iteration, we handle a series of constrained least squares subproblem w.r.t. $\boldsymbol{U}_m$ with an updated weight ($w_{m,j}$) associated with $\widehat{\boldsymbol{R}}_{m,j}$ indicating its reliability; i.e.,

$$\begin{aligned} w_{m,j} &\leftarrow \left(\|\widehat{\boldsymbol{R}}_{m,j} - \widehat{\boldsymbol{U}}_m\widehat{\boldsymbol{U}}_j^\top\|_{\rm F}^2 + \xi\right)^{-\frac{1}{2}}, \\ \widehat{\boldsymbol{U}}_m &\leftarrow \arg\min_{\|\boldsymbol{U}_m\|_{\rm F}\le D}\sum_{j\in\mathcal{S}_{m,j}} w_{m,j}\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m\widehat{\boldsymbol{U}}_j^\top\|_{\rm F}^2, \end{aligned} \tag{11}$$

for all $(m,j)\in\boldsymbol{\Omega}$, where $\xi>0$ is a small number to prevent numerical issues. The procedure in (11) is repeatedly carried out until a certain convergence criterion is met. This algorithm is reminiscent of the classic $\ell_2/\ell_1$ mixed norm minimization (Chartrand & Yin, 2008). Note that the subproblems are fairly easy to handle, as they are quadratic programs; see the supplementary material in Sec. B for more details.

### 3.3. Shifted ReLU Empowered SymNMF

Assume that $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{H}^\top$ is observed (after co-occurrence imputation) with no noise. The task of estimating $\boldsymbol{A}_m$ for all $m$ and $\boldsymbol{D}$ boils down to estimating $\boldsymbol{H}$ from $\boldsymbol{X}$, i.e., a SymNMF problem, as the $\boldsymbol{A}_m$'s can be "extracted" from $\boldsymbol{H}$ easily (cf. Proposition 1). The work in (Huang et al., 2014) offered a simple algorithm for estimating $\boldsymbol{H}\ge\boldsymbol{0}$. Taking the square root decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$, one can see that $\boldsymbol{U} = \boldsymbol{H}\boldsymbol{Q}^\top$ with an orthogonal $\boldsymbol{Q}\in\mathbb{R}^{K\times K}$. It was shown in (Huang et al., 2014) that in the noiseless case, solving the following problem is equivalent to factoring $\boldsymbol{X}$ to $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{H}^\top$ with $\boldsymbol{H}\ge\boldsymbol{0}$:

$$\underset{\boldsymbol{H},\boldsymbol{Q}}{\text{minimize}}\ \ \|\boldsymbol{H} - \boldsymbol{U}\boldsymbol{Q}\|_{\rm F}^2 \tag{12a}$$

$$\text{subject to}\ \ \boldsymbol{H}\ge\boldsymbol{0},\ \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}. \tag{12b}$$

The work in (Huang et al., 2014) proposed an alternating optimization algorithm for handling (12). The algorithm is effective, but it is unclear if it converges to the ground-truth $\boldsymbol{H}$—even without noise. To establish convergence assurances, we propose a simple tweak of the algorithm in (Huang et al., 2014) as follows:

$$\boldsymbol{H}_{(t+1)} \leftarrow \mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{U}\boldsymbol{Q}_{(t)}\right) \tag{13a}$$

$$\boldsymbol{W}_{(t+1)}\boldsymbol{\Sigma}_{(t+1)}\boldsymbol{V}_{(t+1)}^\top \leftarrow \mathsf{svd}\left(\boldsymbol{H}_{(t+1)}^\top\boldsymbol{U}\right) \tag{13b}$$

$$\boldsymbol{Q}_{(t+1)} \leftarrow \boldsymbol{V}_{(t+1)}\boldsymbol{W}_{(t+1)}^\top, \tag{13c}$$

where $\mathsf{ReLU}_\alpha(\cdot):\mathbb{R}^{MK\times K}\to\mathbb{R}^{MK\times K}$ is an elementwise shifted *rectified linear activation function* (ReLU) and is defined as

$$[\mathsf{ReLU}_\alpha(\boldsymbol{Z})]_{i,k} = \begin{cases}\boldsymbol{Z}(i,k), & \text{if } \boldsymbol{Z}(i,k)\ge\alpha,\\ 0, & \text{o.w.},\end{cases}$$

where $\alpha\ge 0$. The step in (13a) is orthogonal projection of each element of $\boldsymbol{U}\boldsymbol{Q}_{(t)}$ to $[\alpha_{(t)},+\infty)$. The two steps (13b) and (13c) give the optimal solution to the $\boldsymbol{Q}$-subproblem, which is often referred to as the *Procrustes projection*. The key difference between our algorithm and the original version in (Huang et al., 2014) is that we use a shifted ReLU function (with a pre-defined sequence $\{\alpha_{(t)}\}$) for the $\boldsymbol{H}$ update, while (Huang et al., 2014) always uses $\alpha_{(t)}=0$. The modification is simple, yet it allows us to offer desirable convergence guarantees. To proceed, we make the following assumption on $\boldsymbol{H}$:

**Assumption 1** *The nonnegative factor* $\boldsymbol{H}\in\mathbb{R}_+^{MK\times K}$ *satisfies: (i)* $\mathrm{rank}(\boldsymbol{H}) = K$ *and* $\|\boldsymbol{H}\|_{\rm F} = \sigma$; *(ii)* $\frac{\|\boldsymbol{H}(j,:)\boldsymbol{\Theta}\|_2^2}{\|\boldsymbol{H}\boldsymbol{\Theta}\|_{\rm F}^2}\le\zeta,\ \forall j,\ \forall\boldsymbol{\Theta}\in\mathbb{R}^{K\times K}$; *(iii) the locations of the nonzero elements of* $\boldsymbol{H}$ *are uniformly distributed over* $[MK]\times[K]$, *and the set* $\boldsymbol{\Delta} = \{(j,k):[\boldsymbol{H}]_{j,k}>0\}$ *has the following cardinality bound*

$$|\boldsymbol{\Delta}| = O\left(\frac{MK\gamma_0^2}{(1+MK\zeta)\sigma^4}\right); \tag{14}$$

*and (iv)* $0<\gamma_0\le\min_{1\le k\le K}\{\beta_k^2-\beta_{k+1}^2\}$, *where* $\beta_k$ *is the* $k$*th singular value of* $\boldsymbol{H}$ *and* $\beta_{K+1}=0$.

Assumption (ii) means that the energy of the range space of $\boldsymbol{H}$ is well spread over its rows. Assumption (iii) means that the nonzero support of $\boldsymbol{H}$ is not too dense. This reflects

the fact that sparsity of the latent factors is often favorable in NMF problems, for both enhancing model identifiability and accelerating computation (Fu et al., 2019; Huang & Sidiropoulos, 2014; Huang et al., 2014). Assumption (iv) means that $\boldsymbol{H}$'s singular values are sufficiently different, which is often useful in characterizing SVD-based operations when noise is present [cf. Eq. (13b)]. With these assumptions, we show the following theorem:

**Theorem 3** *Under Assumption 1, consider* $\widehat{\boldsymbol{U}} = \boldsymbol{H}\boldsymbol{Q}^\top + \boldsymbol{N}$, *where* $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ *is orthogonal, and apply* (13). *Denote* $\nu = \|\boldsymbol{N}\|_{\rm F}$, $h_{(t)} = \|\boldsymbol{H}_{(t)} - \boldsymbol{H}\boldsymbol{\Pi}\|_{\rm F}^2$ *and* $q_{(t)} = \|\boldsymbol{Q}_{(t)} - \boldsymbol{Q}\boldsymbol{\Pi}\|_{\rm F}^2$, *where* $\boldsymbol{\Pi}$ *is any permutation matrix. Suppose that* $\nu \le \sigma \min\{(1-\rho)\sqrt{\eta}q_{(0)}, 1\}$ *for* $\rho := O(K\eta\sigma^4/\gamma_0^2) \in (0,1)$, *where* $\eta = (|\boldsymbol{\Delta}|/MK^2)(1+MK\zeta)$, *and that*

$$2\sigma q_{(0)} + 2\nu < \min_{(j,k)\in\boldsymbol{\Delta}} [\boldsymbol{H}]_{j,k}. \tag{15}$$

*Then, there exists* $\alpha_{(t)} = \alpha > 0$ *such that with probability of at least* $1-\delta$, *the following holds:*

$$q_{(t)} \le \rho q_{(t-1)} + O\left(K\sigma^2\nu^2/\gamma_0^2\right), \tag{16a}$$
$$h_{(t)} \le 2\eta\sigma^2 q_{(t-1)} + 2\nu^2, \tag{16b}$$

*where* $\delta = 2\exp\left(-2|\boldsymbol{\Delta}|/K^2\left(1-\frac{|\boldsymbol{\Delta}|-1}{MK^2}\right)\right)$.

The proof is relegated to the supplementary material in Sec. F. Theorem 3 can be understood as that the solution sequence produced by the algorithm in (13) converges *linearly* to neighborhoods of the ground-truth latent factors (up to a column permutation ambiguity)—and the neighborhoods have zero volumes if noise is absent. Specifically, Eq. (16a) means that, with high probability, the estimation error of $\boldsymbol{Q}$ decreases by a factor of $\rho$ after each iteration—which corresponds to a linear (geometric) rate. Consequently, by Eq. (16b), the estimation error of $\boldsymbol{H}$ also declines in the same rate.

The theorem is also consistent with some long-existing empirical observations from the NMF literature. For example, the parameter $\eta$ is proportional to the number of nonzero elements in the latent factor $\boldsymbol{H}$. Apparently, a sparser $\boldsymbol{H}$ induces a smaller $\eta$, and thus a smaller $\rho$—which means faster convergence. The fact that NMF algorithms in general are in favor of sparser latent factors was previously observed and articulated from multiple perspectives (Gillis, 2012; Huang & Sidiropoulos, 2014; Huang et al., 2014).

A remark is that the convergence result in Theorem 3 holds if the initialization is reasonable [cf. Eq. (15)]. Nevertheless, our experiments show that simply using $\boldsymbol{Q}_{(0)} = \boldsymbol{I}$ works well in practice. We also find that using a diminishing sequence of $\{\alpha_{(t)}\}$ often helps accelerate convergence; see more discussions in the supplementary material in Sec. C.1.2.

Convergence analysis for (Sym)NMF algorithms is in general challenging due to the NP-hardness, even without any noise (Vavasis, 2010). Provable NMF algorithms without relying on restrictive conditions like "separability" (see definition in (Donoho & Stodden, 2003)) are rarely seen in the literature. Notably, the work in (Li & Liang, 2017; Li et al., 2016) also used $\mathsf{ReLU}_\alpha(\cdot)$ for guaranteed NMF—but their algorithms are not for SymNMF and the analyses cannot be applied to our orthogonality-constrained problem.

**Complexity.** The steps in (13a) and (13b) and the Procrustes projection in (13c) both cost $O(MK^3)$ flops. The SVD in (13b) requires $O(K^3)$ flops. Note that in crowdsourcing, $K$ is the number of classes, which is normally small relative to $M$ (the number of annotators). Hence, the algorithm often runs with a competitive speed.

## 4. Experiments

**Baselines.** We denote the proposed robust co-occurrence imputation-assisted SymNMF algorithm as `RobSymNMF` and the designated annotators-based imputation-based SymNMF as `DesSymNMF`. To benchmark our methods, we employ a number of crowdsourcing algorithms, namely, `MultiSPA`, `CNMF` (Ibrahim et al., 2019), `TensorADMM` (Traganitis et al., 2018) `Spectral-D&S` (Zhang et al., 2016), `KOS` (Karger et al., 2013), `EigenRatio` (Dalvi et al., 2013), `GhoshSVD` (Ghosh et al., 2011), and `MinimaxEntropy` (Zhou et al., 2014). We also employ `EM` (Dawid & Skene, 1979) initialized by `majority voting` (denoted as `MV-EM`) as a baseline. Note that `CNMF` is the state-of-the-art, which uses pairwise co-occurrences as our methods do. We also use our proposed methods to initialize `EM` (`RobSymNMF-EM` and `DesSymNMF-EM`). For all the D&S model-based algorithms, we construct an MAP predictor for $y_n$ after the model is learned.

**Synthetic Data Experiments.** The synthetic data experiments are presented in the supplementary material in Sec. C.1.

**UCI Data Experiments.** We consider a number of UCI datasets, namely, "Connect4", "Credit" and "Car". We choose different classifiers from the MATLAB machine learning toolbox, e.g., support vector machines and decision tree; see Sec. C.2 of the supplementary material for details. These classifiers serve as annotators in our experiments. We partition the datasets randomly in every trial, with a training to testing ratio being 1/4—which means that the annotators are not extensively trained. Each classifier (annotator) is then allowed to label a test item with probability $p_m \in (0,1]$.

Tables 1 and 2 show the performance of the algorithms on

*Table 1.* Classification error (%) and runtime (sec.) on the UCI Connect4 dataset ($N = 20,561$, $M = 10$, $K = 3$). The "SymNMF" family are the proposed methods.

| Algorithms | $p_m = 0.3$ | $p_m \in (0.3, 0.5)$, $p_d = 0.8$ | $p_m \in (0.5, 0.7)$, $p_d = 0.8$ | Time(s) |
|---|---|---|---|---|
| RobSymNMF | **33.26** | 33.06 | 32.16 | 0.142 |
| RobSymNMF-EM | 34.27 | 33.20 | 32.11 | 0.191 |
| RobSymNMF ($w_{m,j} = 1$) | **33.14** | 34.60 | 33.91 | 0.132 |
| DesSymNMF | 33.45 | **32.18** | **31.42** | 0.061 |
| DesSymNMF-EM | 33.94 | **32.50** | **31.40** | 0.128 |
| SymNMF (w/o imput.) | 34.87 | 35.71 | 32.00 | 0.052 |
| MultiSPA | 47.78 | 42.24 | 49.54 | 0.020 |
| CNMF | 36.26 | 39.55 | 34.70 | 4.741 |
| TensorADMM | 36.20 | 34.34 | 35.18 | 5.183 |
| Spectral-D&S | 64.28 | 66.95 | 71.97 | 20.388 |
| MV-EM | 34.14 | 34.17 | 34.19 | 0.107 |
| MinimaxEntropy | 36.20 | 36.17 | 35.46 | 27.454 |
| KOS | 54.55 | 43.21 | 39.41 | 12.798 |
| Majority Voting | 37.76 | 36.88 | 36.75 | - |

*Table 2.* Classification error (%) and runtime (sec.) on the UCI Credit dataset ($N = 540$, $M = 10$, $K = 2$). The "SymNMF" family are the proposed methods.

| Algorithms | $p_m = 0.3$ | $p_m \in (0.3, 0.5)$, $p_d = 0.8$ | $p_m \in (0.5, 0.7)$, $p_d = 0.8$ | Time(s) |
|---|---|---|---|---|
| RobSymNMF | **16.31** | **13.99** | 13.74 | 0.152 |
| RobSymNMF-EM | 16.76 | 13.96 | 14.06 | 0.160 |
| RobSymNMF ($w_{m,j} = 1$) | **16.32** | **13.99** | **13.72** | 0.062 |
| DesSymNMF | 16.37 | **13.83** | **13.67** | 0.052 |
| DesSymNMF-EM | 16.80 | 14.07 | 13.77 | 0.059 |
| SymNMF (w/o imput.) | 16.51 | 13.94 | 13.85 | 0.039 |
| MultiSPA | 16.74 | 14.28 | 14.60 | 0.003 |
| CNMF | 16.74 | 14.24 | 14.40 | 3.273 |
| TensorADMM | 16.70 | 14.31 | 13.87 | 3.405 |
| Spectral-D&S | 16.98 | 14.24 | 14.00 | 1.790 |
| MV-EM | 44.54 | 26.20 | 14.00 | 0.007 |
| MinimaxEntropy | 17.50 | 17.00 | 16.78 | 0.728 |
| KOS | 17.28 | 14.22 | 14.89 | 0.009 |
| GhoshSVD | 17.07 | 14.76 | 14.80 | 0.009 |
| EigenRatio | 17.17 | 14.43 | 14.44 | 0.003 |
| Majority Voting | 18.22 | 15.95 | 14.83 | - |

Connect4 and Credit, respectively. In the first column of the tables, $p_m$ is fixed for all $M$ annotators. In the second and third columns, we designate two annotators $\ell$ and $r$, and let them label the data items with higher probabilities (i.e., $p_d$). This way, the designated annotators can co-label items with many other annotators—which can help impute missing co-occurrences using (7)-(8). The designated annotators $\ell$ and $r$ are chosen from the $M$ annotators randomly in each trial. The probability $p_m$ is also randomly chosen from a pre-specified range as indicated in the tables. We use this setting to simulate realistic scenarios in crowdsourcing where incomplete, noisy, and unbalanced labels are present. The results are averaged from 20 trials.

From Tables 1 and 2, one can observe that the proposed methods show promising classification performance in all cases. The proposed methods exhibit clear improvements upon the CNMF—especially in the more challenging case in Table 1. The proposed methods also outperform the the third-order statistics-based ones (TensorADMM and Spectral-D&S) under most settings, articulating the advantages of using second-order statistics. In terms of the runtime performance, the proposed SymNMF family are also about 20 to 50 times faster compared to CNMF in these two tables. There are 10% of co-occurrences missing in the cases corresponding to the first columns of Tables 1 and 2. DesSymNMF using (7)-(8) is able to impute all the missing ones, although we did not assign any designated annotator. In both tables, RobSymNMF slightly (but consistently) outperforms DesSymNMF when there is no designated annotators, showing some advantages in such cases. In the above experiments, our robust imputation algorithm in (11) offers labeling errors that are smaller than or equal to its non-robust version (with $w_{m,j} = 1$) in 5 out of 6 settings.

Table 3 presents the performance of the algorithms on the Car dataset under different proportions of missing co-occurrences; see Sec. C.2 of the supplementary material for the details of generating such cases. In this experiment, we do not assign designated annotators. If $\boldsymbol{R}_{m,n}$ cannot be completed by observed co-occurrences using (7)-(8), we leave it as an all-zero block. Using (7) and (8), DesSymNMF still improves the missing proportions to 17%, 9% and 0% for the columns from left to right, respectively. One can see that the proposed method largely outperforms the baselines, especially in the cases where 70% of the $\boldsymbol{R}_{m,j}$'s are not observed. However, CNMF is not able to produce competitive results in this experiment.

*Table 3.* Classification error (%) and runtime (sec.) on the UCI Car dataset ($N = 1,352$, $M = 10$, $K = 4$). The "SymNMF" family are the proposed methods.

| Algorithms | Miss = 70% | Miss = 50% | Miss = 30% | Time (s) |
|---|---|---|---|---|
| RobSymNMF | **24.01** | **23.17** | **22.05** | 0.108 |
| RobSymNMF-EM | 24.93 | 23.71 | **22.03** | 0.123 |
| RobSymNMF ($w_{m,j} = 1$) | **24.01** | **23.40** | 22.16 | 0.100 |
| DesSymNMF | 24.50 | 23.41 | 23.00 | 0.048 |
| DesSymNMF-EM | 24.91 | 24.59 | 23.45 | 0.060 |
| SymNMF (w/o imput.) | 24.43 | 24.03 | 24.40 | 0.031 |
| MultiSPA | 47.12 | 47.14 | 33.84 | 0.002 |
| CNMF | 43.65 | 41.49 | 30.55 | 3.666 |
| TensorADMM | 36.67 | 39.32 | 37.38 | 4.900 |
| Spectral-D&S | 31.20 | 29.67 | 29.14 | 47.800 |
| MV-EM | 30.27 | 29.96 | 29.65 | 0.013 |
| MinimaxEntropy | 28.22 | 25.73 | 24.68 | 12.664 |
| KOS | 48.87 | 49.87 | 41.83 | 0.104 |
| Majority Voting | 43.88 | 43.08 | 42.40 | - |

**AMT Data Experiments.** We also evaluate the algorithms using various AMT datasets, namely "Bluebird", "Dog", "RTE" and "TREC", which are annotated by human annotators. The AMT datasets are more challenging, in the sense that we have no control for annotation acquisition and no designated annotators are available. Similar as before, for DesSymNMF, we leave the co-occurrences that cannot be recovered by (7)-(8) as all-zero blocks. In the AMT experiments, we include two additional baselines based on tensor completion, namely, PG-TAC (Zhou & He, 2016) and CRIA$_V$ (Li & Jiang, 2018)—both of which reported good performance over AMT datasets.

*Table 4.* Classification error (%) and runtime (sec.) on the AMT datasets "Bluebird" and "Dog".The "SymNMF" family are the proposed methods.

| Algorithms | Bluebird ($N = 108$, $M = 39$, $K = 2$) | | Dog ($N = 807$, $M = 52$, $K = 4$) | |
|---|---|---|---|---|
| | **Error** (%) | **Time** (s) | **Error** (%) | **Time** (s) |
| RobSymNMF | 11.11 | 0.72 | **16.10** | 0.41 |
| RobSymNMF-EM | 11.11 | 0.79 | **15.86** | 0.48 |
| RobSymNMF ($w_{m,j} = 1$) | 11.11 | 0.38 | **16.10** | 0.38 |
| DesSymNMF | **10.18** | 0.15 | 16.35 | 0.11 |
| DesSymNMF-EM | **10.18** | 0.19 | **15.86** | 0.16 |
| SymNMF (w/o imput.) | **10.18** | 0.12 | 16.72 | 0.10 |
| MultiSPA | 13.88 | 0.10 | 17.96 | 0.09 |
| CNMF | 11.11 | 6.76 | **15.86** | 17.14 |
| TensorADMM | 12.03 | 85.56 | 18.01 | 613.93 |
| Spectral-D&S | 12.03 | 1.97 | 17.84 | 43.88 |
| MV-EM | 12.03 | 0.02 | **15.86** | 0.06 |
| MinimaxEntropy | **8.33** | 3.43 | 16.23 | 4.6 |
| KOS | 11.11 | 0.11 | 31.84 | 0.17 |
| GhoshSVD | 27.77 | 0.02 | N/A | N/A |
| EigenRatio | 27.77 | 0.03 | N/A | N/A |
| PG-TAC | 24.07 | 0.04 | 18.21 | 21.11 |
| CRIA$_V$ | 24.07 | 0.05 | 17.10 | 18.48 |
| Majority Voting | 21.29 | N/A | 17.91 | N/A |

*Table 5.* Classification error (%) and runtime (sec.) on the AMT datasets "RTE" and "TREC". The "SymNMF" family are the proposed methods.

| Algorithms | RTE ($N = 800$, $M = 164$, $K = 2$) | | TREC ($N = 19,033$, $M = 762$, $K = 2$) | |
|---|---|---|---|---|
| | **Error** (%) | **Time** (s) | **Error** (%) | **Time** (s) |
| RobSymNMF | **7.25** | 2.31 | 30.68 | 64.99 |
| RobSymNMF-EM | **7.12** | 2.4 | 29.62 | 67.39 |
| RobSymNMF ($w_{m,j} = 1$) | 7.37 | 1.35 | 33.23 | 62.33 |
| DesSymNMF | 13.87 | 3.32 | 36.75 | 71.31 |
| DesSymNMF-EM | **7.25** | 3.43 | **29.36** | 72.13 |
| SymNMF (w/o imput.) | 48.75 | 0.23 | 35.47 | 57.60 |
| MultiSPA | 8.37 | 0.18 | 31.56 | 51.34 |
| CNMF | **7.12** | 18.12 | 29.84 | 536.86 |
| TensorADMM | N/A | N/A | N/A | N/A |
| Spectral-D&S | **7.12** | 6.34 | **29.58** | 919.98 |
| MV-EM | **7.25** | 0.09 | 30.02 | 3.12 |
| MinimaxEntropy | 7.5 | 6.4 | 30.89 | 356.32 |
| KOS | 39.75 | 0.07 | 51.95 | 8.53 |
| GhoshSVD | 49.12 | 0.06 | 43.03 | 7.18 |
| EigenRatio | 9.01 | 0.07 | 43.95 | 1.87 |
| PG-TAC | 8.12 | 50.41 | 33.89 | 917.21 |
| CRIA$_V$ | 9.37 | 49.04 | 34.59 | 900.34 |
| Majority Voting | 10.31 | N/A | 34.85 | N/A |

Table 4 and 5 present the evaluation results over the AMT datasets. The TensorADMM algorithm could not run with large $M$ due to scalablity issues. The results are consistent with those observed in the UCI experiments. The proposed methods' labeling accuracy is either comparable with or better than that of CNMF, but is order-of-magnitude faster. The proposed methods are also observed to most effectively initialize the EM algorithm (Dawid & Skene, 1979). An observation is that there are 2.5%, 14.0%, 90.68%, and 96.57% of the pairwise co-occurrences missing in Bluebird, Dog, RTE and TREC, respectively. DesSymNMF is able to bring down the missing proportions to 0.00%. 11.34%, 50.15%, and 92.18%, respectively. The DesSymNMF imputation can sometimes improve the final accuracy significantly; see the Dog and RTE columns. In addition, our robust imputation criterion (9) and the algorithm in (11) often exhibit visible improvements upon the equally weighted (non-robust) version, as in the UCI case.

**Comparison with Deep Learning-based Methods.** We present an additional experiment and compare the proposed approaches with two deep learning (DL)-based crowdsourcing methods in (Rodrigues & Pereira, 2018). The details can be found in the supplementary material in Sec. C.3.

## 5. Conclusion

We proposed a D&S model identification-based crowdsourcing method that uses sample-efficient pairwise co-occurrences of annotator responses. We advocated a SymNMF-based framework that offers strong identifiability of the D&S model under reasonable conditions. To realize the SymNMF framework, we proposed two lightweight algorithms for provably imputing missing co-occurrences when the annotations are incomplete. We also proposed a computationally economical SymNMF algorithm, and analyzed its convergence properties. We tested the framework on UCI and AMT data and observed promising performance. The proposed algorithms are typically order-of-magnitude faster than other high-performance baselines.

## 6. Acknowledgement

This work is supported in part by the National Science Foundation under Project NSF IIS-2007836 and the Army Research Office under Project ARO W911NF-19-1-0247.

## References

Buhrmester, M. D., Kwang, T., and Gosling, S. Amazon's mechanical turk. *Perspectives on Psychological Science*, 6:3–5, 2011.

Candés, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(3), 2011.

Chartrand, R. and Yin, W. Iteratively reweighted algorithms for compressive sensing. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 3869 –3872, 2008.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of International Conference on World Wide Web*, pp. 285–294, 2013.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pp. 20–28, 1979.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts?

In *Advances in neural information processing systems*, volume 16, 2003.

Fu, X., Ma, W.-K., Huang, K., and Sidiropoulos, N. D. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Trans. Signal Process.*, 63(9):2306–2320, May 2015.

Fu, X., Huang, K., Yang, B., Ma, W.-K., and Sidiropoulos, N. D. Robust volume minimization-based matrix factorization for remote sensing and document clustering. *IEEE Trans. Signal Process.*, 64(23):6254–6268, 2016.

Fu, X., Huang, K., and Sidiropoulos, N. D. On identifiability of nonnegative matrix factorization. *IEEE Signal Process. Lett.*, 25(3):328–332, 2018.

Fu, X., Huang, K., Sidiropoulos, N. D., and Ma, W.-K. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Process. Mag.*, 36(2):59–80, 2019.

Fu, X., Ibrahim, S., Wai, H.-T., Gao, C., and Huang, K. Block-randomized stochastic proximal gradient for low-rank tensor factorization. *IEEE Trans. Signal Process.*, 68:2170–2185, 2020a.

Fu, X., Vervliet, N., De Lathauwer, L., Huang, K., and Gillis, N. Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective. *IEEE Signal Process. Mag.*, 37 (5):78–94, 2020b.

Ghosh, A., Kale, S., and McAfee, P. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM conference on Electronic commerce*, pp. 167–176, 2011.

Gillis, N. Sparse and unique nonnegative matrix factorization through data preprocessing. *The Journal of Machine Learning Research*, 13(1):3349–3386, 2012.

Gillis, N. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, 2020.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Han, Y., Jiao, J., and Weissman, T. Minimax estimation of discrete distributions under $l_1$ loss. *IEEE Trans. Inf. Theory*, 61(11):6343–6354, 2015.

Huang, K. and Sidiropoulos, N. Putting nonnegative matrix factorization to the test: a tutorial derivation of pertinent Cramer-Rao bounds and performance benchmarking. *IEEE Signal Process. Mag.*, 31(3):76–86, 2014.

Huang, K., Sidiropoulos, N., and Swami, A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Process.*, 62(1):211–224, 2014.

Ibrahim, S., Fu, X., Kargas, N., and Huang, K. Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7847–7857, 2019.

Karger, D., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1953–1961, 2011a.

Karger, D. R., Oh, S., and Shah, D. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Annual Allerton Conference on Communication, Control, and Computing*, pp. 284–291, 2011b.

Karger, D. R., Oh, S., and Shah, D. Efficient crowdsourcing for multi-class labeling. *ACM Sigmetrics Performance Evaluation Review*, 41(1):81–92, 2013.

Karger, D. R., Oh, S., and Shah, D. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 453–456, 2008.

Li, H. *Theoretical analysis and efficient algorithms for crowdsourcing*. PhD thesis, UC Berkeley, 2015.

Li, H. and Yu, B. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv:1411.4086*, 2014.

Li, S.-Y. and Jiang, Y. Multi-label crowdsourcing learning with incomplete annotations. In *PRICAI 2018: Trends in Artificial Intelligence*, pp. 232–245, 2018.

Li, Y. and Liang, Y. Provable alternating gradient descent for non-negative matrix factorization with strong correlations. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 2062–2070. PMLR, 06–11 Aug 2017.

Li, Y., Liang, Y., and Risteski, A. Recovery guarantee of non-negative matrix factorization via alternating updates. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4987–4995, 2016.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, volume 25, pp. 692–700, 2012.

Ma, Y., Olshevsky, A., Szepesvari, C., and Saligrama, V. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. In *Proceedings of International Conference on Machine Learning*, volume 80, pp. 3335–3344, 2018.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1–21, 2015.

Nie, F., Yuan, J., and Huang, H. Optimal mean robust principal component analysis. In *International Conference on Machine Learning*, pp. 1062–1070, 2014.

Rodrigues, F. and Pereira, F. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Salk, C. F., Sturn, T., See, L., and Fritz, S. Limitations of majority agreement in crowdsourced image interpretation. *Transactions in GIS*, 21(2):207–223, 2017.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

Sun, R. and Luo, Z.-Q. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory*, 62 (11):6535–6579, 2016.

Traganitis, P. A., Pages-Zamora, A., and Giannakis, G. B. Blind multiclass ensemble classification. *IEEE Trans. Signal Process.*, 66(18):4737–4752, 2018.

Vakharia, D. and Lease, M. Beyond AMT: an analysis of crowd work platforms. *Computing Research Repository*, 2013.

Vavasis, S. A. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

Wazny, K. "crowdsourcing" ten years in: A review. *Journal of Global Health*, 7:020602, 2017.

Welinder, P., Branson, S., Perona, P., and Belongie, S. J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pp. 2424–2432, 2010.

Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pp. 2035–2043. 2009.

Xu, H., Caramanis, C., and Sanghavi, S. Robust PCA via outlier pursuit. *IEEE Trans. Inf. Theory*, 58(5):3047–3064, 2012.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2195–2203. 2012.

Zhou, D., Liu, Q., Platt, J., and Meek, C. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of International Conference on Machine Learning*, volume 32, pp. 262–270, 2014.

Zhou, D., Liu, Q., Platt, J. C., Meek, C., and Shah, N. B. Regularized minimax conditional entropy for crowdsourcing. *Computing Research Repository*, 2015.

Zhou, Y. and He, J. Crowdsourcing via tensor augmentation and completion. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2435–2441, 2016.