Crowdsourcing via Pairwise Co-occurrences: Identifiability and Algorithms

Shahana Ibrahim 1, Xiao Fu 1, Nikos Kargas 2, and Kejun Huang 1

1School of Elect. Eng. & Computer Sci., Oregon State University, Corvallis, USA 2Department of Elect. & Computer Eng., University of Minnesota, Minneapolis, MN USA

Department of Computing & Info. Sci. & Eng., University of Florida, Gainesville, FL, USA

# Our Approach

1. We first normalize the columns of $Z_n$ to get $\mathbf{Z}_n = A_n \mathcal{T}_n$, where $\mathcal{T}_n$ is now normalized and $A_n$ is column normalized by definition.

2. Now, let us assume there exists some rows in $\mathcal{T}_n$, with index $\ell \in \{1, \ldots, q\}$, such that:

$$A_{n}(\ell, \bullet) = 1.$$

3. This condition is known as separability in Nonnegative Matrix Factorization [Guillén & Vaness, 2014].

4. Under the separability assumption, our task boils down to identifying $\ell$, an index selection problem which can be achieved by using Successive Projection Algorithm (SPA) [Kato et al., 2015].

5. We repeat this index identification procedure to identify all $\ell_n$, and name our approach MultiSPA.

# Enhanced Identifiability

1. If every class has an annotator who can perfectly identify class $\ell$, then $\mathcal{T}_n(\ell, \bullet)$ may be satisfied.

2. Satisfying $\mathcal{T}_n(\ell, \bullet) = 1$ may be too ideal.

3. In practice, the annotations may not be perfect for any class, but can be reasonably good for some class.

### Theorem 1

Assume that $\mathcal{T}_n(\ell, \bullet)$ is not an exactly identifiable matrix, SPA can still identify the indices.

### Theorem 2

Let $p > q$ be such that, and assume that the rows of $\mathcal{T}_n$ are generated within the $\mathcal{T}_n$ probability simplex uniformly at random. If the number of annotators satisfies $n \geq \Omega(\max\{p, q\})$, then, with probability greater than or equal to $1 - p$ they exist rows of $\mathcal{T}_n$ indexed by $q$, such that $\mathcal{T}_n(q, \bullet)$ is a sparse vector identifying the indices, i.e., $A_{n}(q, \bullet) = 1$.

## Do we favour more annotators?

### Theorem 3

Assume that $\mathcal{T}_n(\ell, \bullet) = 1$ for all $\ell = 1, \ldots, M$, and that there exist two subsets of the annotator indexed by $P_1$ and $P_2$, where $P_1 \cap P_2 = \emptyset$, and $P_1 \cup P_2 \subseteq \{1, \ldots, M\}$. Suppose that from $P_1$ and $P_2$ the following two matrices can be constructed: $H_1 = [A_{1}(P_1, \bullet)]$, $H_2 = [A_{2}(P_2, \bullet)]$, where $m_1 \in P_1$, and $m_2 \in P_2$. Furthermore, assume that $A_{1}(m_1, \bullet)$ and $A_{2}(m_2, \bullet)$ are sufficiently sparse, there exists an $A_k$ available, and for each $m \in P_1 \cup P_2$ there exists a $A_k$ available, where $m \in P_1 \cup P_2$. Then, solving the coupled matrix decomposition problem returns $A_{n}(\ell, \bullet)$ up to identical column permutations.

### Alternating Optimization KL Algorithm

1. We use KL divergence as fitting criterion in the coupled matrix decomposition problem.

2. The scalable algorithm MultiSPA can be used as initialization to the KL algorithm and is named as MultiSPA-KL.

### Amazon Mechanical Turk Experiment Results

1. The datasets annotated by Amazon Mechanical Turk (https://www.mturk.com) (AMT) workers are used here.

---

**References**


---

**Notes:**

- Minmax-entropy
- MV-D&S
- TensorADMM
- MultiSPA-D&S
- MultiSPA-KL
- MultiSPA