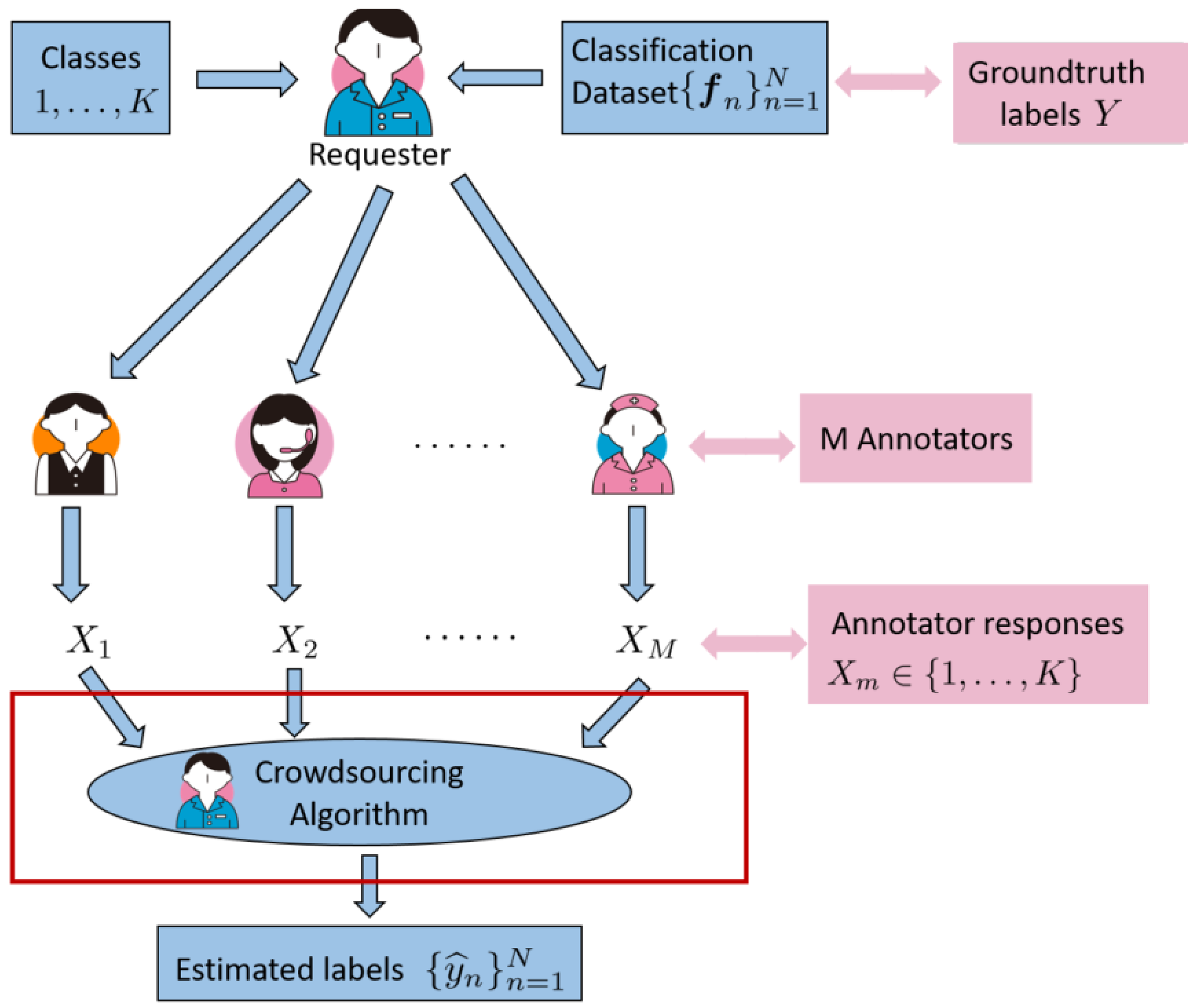
Crowdsourcing via Pairwise Co-occurrences: Identifiability and Algorithms

Shahana Ibrahim[†], Xiao Fu[†], Nikos Kargas[‡], and Kejun Huang[⋆]

[†]School of EECS, Oregon State University, Corvallis, OR, USA [‡]Department of ECE, University of Minnesota, Minneapolis, MN, USA
[⋆]Department of CISE, University of Florida, Gainesville, FL, USA

Data Labeling and Crowdsourcing

- Massive labeled data is a key performance booster of deep networks.
- **Crowdsourcing** is widely used for data labeling.



Dawid-Skene Model

- The **confusion matrix** $\boldsymbol{A}_m \in \mathbb{R}^{K\times K}$ for each annotator m and the **prior probability vector** $\boldsymbol{d} \in \mathbb{R}^K$ are the Dawid-Skene model parameters.

$$\boldsymbol{A}_m(k_m,k) := \Pr(X_m = k_m | Y = k),$$
$$\boldsymbol{d}(k) := \Pr(Y = k)$$

- **The goal is to estimate $\boldsymbol{A}_m$ for $m = 1,\ldots,M$ and $\boldsymbol{d}$.**

Prior Art

- **Dawid-Skene Model** [Dawid & Skene, 1979]:
  - Proposed expectation maximization (EM) algorithm for ML estimation.
  - Widely used, but model identifiability is unclear.
- **Spectral Method** [Zhang et al., 2014]:
  - Established identifiability using orthogonal and symmetric tensor decomposition.
  - Employed third-order co-occurrences of responses; may have high sample complexity.

Pairwise Co-occurrences of Annotator Responses

- The joint PMF of any two annotator responses,

$$\boldsymbol{R}_{m,\ell}(k_m,k_\ell) = \sum_{k=1}^{K} \underbrace{\Pr(Y=k)}_{\boldsymbol{d}(k)} \underbrace{\Pr(X_m = k_m|Y=k)}_{\boldsymbol{A}_m(k_m,k)} \underbrace{\Pr(X_\ell = k_\ell|Y=k)}_{\boldsymbol{A}_\ell(k_\ell,k)},$$
$$\Longrightarrow \boldsymbol{R}_{m,\ell} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_\ell^\top \in \mathbb{R}^{K\times K}, \ \boldsymbol{D} = \mathrm{Diag}(\boldsymbol{d}).$$

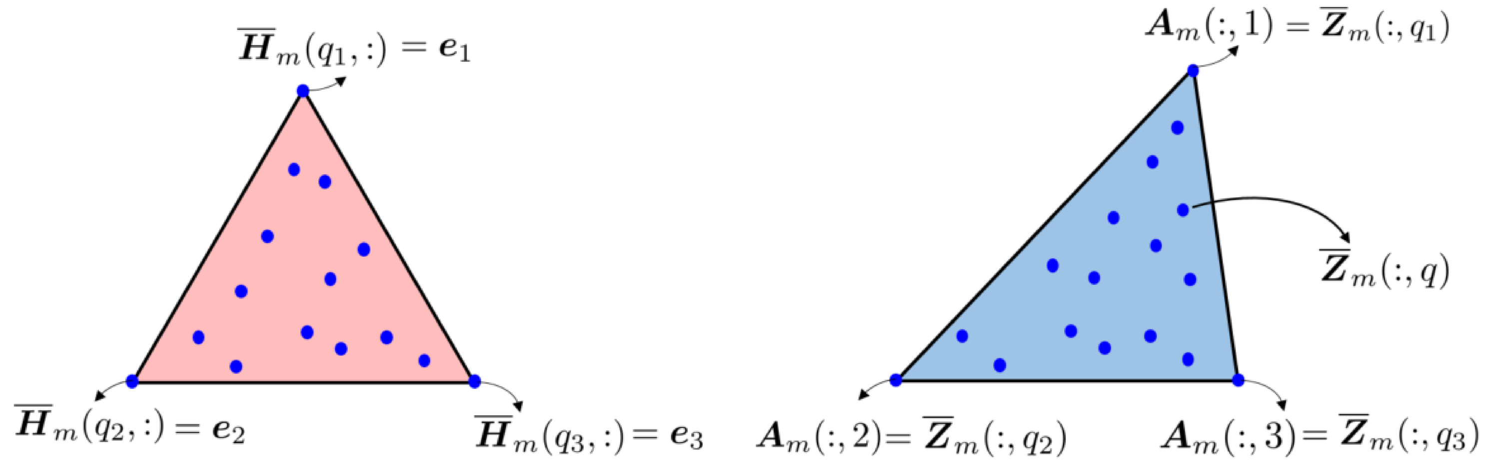
- $\boldsymbol{R}_{m,\ell}$'s can be estimated via sample averaging.
- $\boldsymbol{R}_{m,\ell}$'s are second-order statistics; easier to estimate than third-order ones.

Proposed Approach

- Consider an annotator m who co-labels with annotators $m_1,\ldots,m_{T(m)}$,

$$\boldsymbol{Z}_m = [\boldsymbol{R}_{m,m_1}, \boldsymbol{R}_{m,m_2}, \ldots, \boldsymbol{R}_{m,m_{T(m)}}] = \boldsymbol{A}_m \underbrace{[\boldsymbol{D}\boldsymbol{A}_{m_1}^\top, \ldots, \boldsymbol{D}\boldsymbol{A}_{T(m)}^\top]}_{\boldsymbol{H}_m^\top}.$$

- ℓ_1-normalize the columns of $\boldsymbol{Z}_m$ to get $\overline{\boldsymbol{Z}}_m = \boldsymbol{A}_m \overline{\boldsymbol{H}}_m^\top$ where $\overline{\boldsymbol{H}}_m^\top$ is row normalized.
- Assume that there exits an index set $\Lambda_q = \{q_1,\ldots,q_K\}$ such that $\overline{\boldsymbol{H}}_m(\Lambda_q,:) = \boldsymbol{I}_K$ (known as **seperability**) [Donoho & Stodden, 2003].



- Estimating $\boldsymbol{A}_m$ boils down to identifyting index set Λ_q which can be achieved by **successive projection algorithm (SPA)** [Araújo et al. 2001].
- Index identification via SPA is repeated for every $\boldsymbol{A}_m$ (named as `MultiSPA`).

Model Identifiability

- If each class k has an annotator who can perfectly identify class k, then $\overline{\boldsymbol{H}}_m(\Lambda_q,:) = \boldsymbol{I}_K$ can be satisfied.

$$\boldsymbol{Z}_m = \boldsymbol{A}_m \underbrace{\boldsymbol{D}[\boldsymbol{A}_{m_1}^\top, \ldots, \overset{\boldsymbol{e}_1}{\boldsymbol{A}_{m_{e1}}^\top}, \ldots, \overset{\boldsymbol{e}_2}{\boldsymbol{A}_{m_{e2}}^\top}, \ldots, \overset{\boldsymbol{e}_3}{\boldsymbol{A}_{m_{e3}}^\top}, \ldots, \boldsymbol{A}_{T(m)}^\top]}_{\boldsymbol{H}_m^\top}$$

Theorem 1: Assume that annotators m and t co-label at least S samples $\forall t \in \{m_1,\ldots,m_{T(m)}\}$, Also assume that the constructed $\widehat{\boldsymbol{Z}}_m$ satisfies $\|\widehat{\boldsymbol{Z}}_m(:,l)\|_1 \geq \eta, \forall l \in \{1,\ldots KT(m)\}$, where $\eta \in (0,1]$. Suppose that for every class index $k \in \{1,\ldots,K\}$, there exists an annotator $m_{t(k)} \in \{m_1,\ldots,m_{T(m)}\}$ such that
$$\Pr(X_{m_{t(k)}} = k|Y=k) \geq (1-\epsilon)\sum_{j=1}^{K} \Pr(X_{m_{t(k)}} = k|Y=j), \ \epsilon \in [0,1]$$
Then, if $\epsilon \leq \mathcal{O}\left(\max\left(K^{-1}\kappa^{-3}(\boldsymbol{A}_m), \sqrt{\ln(1/\delta)}(\sigma_{\max}(\boldsymbol{A}_m)\sqrt{S}\eta)^{-1}\right)\right)$, with probability greater than $1-\delta$, the SPA algorithm can estimate an $\widehat{\boldsymbol{A}}_m$ from $\boldsymbol{Z}_m = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{H}_m^\top$ with the estimation error bounded by $\mathcal{O}(\sqrt{K}\kappa^2(\boldsymbol{A}_m)\max(\sigma_{\max}(\boldsymbol{A}_m)\epsilon, \sqrt{\ln(1/\delta)}(\sqrt{S}\eta)^{-1}))$ where $\sigma_{\max}(\boldsymbol{A}_m)$ is the largest singular value of $\boldsymbol{A}_m$, and $\kappa(\boldsymbol{A}_m)$ is the condition number of $\boldsymbol{A}_m$.

- **Implication:** Even if there are no perfect annotators for each class, `MultiSPA` estimates $\boldsymbol{A}_m$.

Do we favour more annotators?

Theorem 2 :Let $\rho > 0, \varepsilon > 0$, and assume that the rows of $\overline{\boldsymbol{H}}_m$ are generated within the $(K-1)$-probability simplex uniformly at random. If $M \geq \Omega\left(\frac{\varepsilon^{-2(K-1)}}{K}\log\left(\frac{K}{\rho}\right)\right)$, then with probability greater than or equal to $1-\rho$, there exists rows of $\overline{\boldsymbol{H}}_m$ indexed by $q_1,\ldots q_K$ such that
$$\|\overline{\boldsymbol{H}}_m(q_k,:) - \boldsymbol{e}_k^\top\|_2 \leq \varepsilon, \ k = 1,\ldots,K.$$

- **Implication:** If more number of annotators are available, there exists high chance for seperability condition.

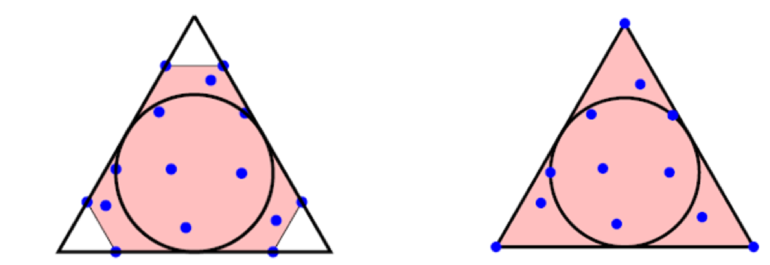
Enhanced Identifiability

- The model can be identified under a relaxed assumption by solving

$$\text{find} \quad \{\boldsymbol{A}_m\}_{m=1}^M, \boldsymbol{D} \qquad (1a)$$
$$\text{subject to} \quad \boldsymbol{R}_{m,\ell} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_\ell^\top, \ \forall m,\ell \in \{1,\ldots,M\} \qquad (1b)$$
$$\boldsymbol{1}^\top\boldsymbol{A}_m = \boldsymbol{1}^\top, \ \boldsymbol{A}_m \geq \boldsymbol{0}, \ \forall m, \ \boldsymbol{1}^\top\boldsymbol{d} = 1, \ \boldsymbol{d} \geq \boldsymbol{0}. \qquad (1c)$$

Theorem 3 : Assume that $\mathrm{rank}(\boldsymbol{D}) = \mathrm{rank}(\boldsymbol{A}_m) = K$ for all $m = 1,\ldots,M$, and that there exist two subsets of the annotator, indexed by $\mathcal{P}_1$ and $\mathcal{P}_2$, where $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$ and $\mathcal{P}_1 \cup \mathcal{P}_2 \subseteq \{1,\ldots,M\}$. Suppose that from $\mathcal{P}_1$ and $\mathcal{P}_2$ the following two matrices can be constructed:
$$\tilde{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{R}_{m_1,\ell_1} & \boldsymbol{R}_{m_1,\ell_2} & \ldots & \boldsymbol{R}_{m_1,\ell_{|\mathcal{P}_2|}} \\ \vdots & \vdots & \ldots & \vdots \\ \boldsymbol{R}_{m_{|\mathcal{P}_1|},\ell_1} & \boldsymbol{R}_{m_1,\ell_2} & \ldots & \boldsymbol{R}_{m_{|\mathcal{P}_2|},\ell_{|\mathcal{P}_2|}} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{A}_{m_1} \\ \vdots \\ \boldsymbol{A}_{m_{|\mathcal{P}_1|}}. \end{bmatrix}}_{\boldsymbol{H}^{(1)}} \boldsymbol{D} \underbrace{[\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_{|\mathcal{P}_2|}}^\top]}_{(\boldsymbol{H}^{(2)})^\top}$$
Denote $\boldsymbol{H}^{(1)} = [\boldsymbol{A}_{m_1}^\top, \ldots, \boldsymbol{A}_{m_{|\mathcal{P}_1|}}^\top]^\top$, $\boldsymbol{H}^{(2)} = [\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_{|\mathcal{P}_2|}}^\top]^\top$, where $m_t \in \mathcal{P}_1$ and $\ell_j \in \mathcal{P}_2$. Furthermore, assume that both $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ are *sufficiently scattered*.Then, solving Problem (1) recovers $\boldsymbol{A}_m$ for $m = 1,\ldots,M$ and $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{d})$ up to identical column permutation.

- Extremely well trained annotators for each class are not required to satisfy sufficiently scattered condition.



Left: Sufficiently scattered $\boldsymbol{H}$; Right: Separable $\boldsymbol{H}$

- Problem (1) is solved by a BCD algorithm with KL divergence as the fitting criterion (used `MultiSPA` as initialization, thus named as `MultiSPA-KL`).

Amazon Mechanical Turk (AMT) Experiment Results

- The datasets annotated by AMT workers are used.

Algorithms	TREC		Bluebird		RTE		Web		Dog	
	(%)Error	(sec)Time	(%)Error	(sec)Time	(%)Error	(sec)Time	(%)Error	(sec)Time	(%)Error	(sec)Time
`MultiSPA`	31.47	50.68	13.88	0.07	8.75	0.28	15.22	0.54	17.09	0.07
`MultiSPA-KL`	**29.23**	536.89	**11.11**	1.94	**7.12**	17.06	**14.58**	12.34	**15.48**	15.88
`MultiSPA-D&S`	29.84	53.14	12.03	0.09	**7.12**	0.32	15.11	0.84	**16.11**	0.12
`Spectral-D&S`	**29.58**	919.98	12.03	1.97	**7.12**	6.40	16.88	179.92	17.84	51.16
`TensorADMM`	N/A	N/A	12.03	2.74	N/A	N/A	N/A	N/A	17.96	603.93
`MV-D&S`	30.02	3.20	12.03	0.02	7.25	0.07	16.02	0.28	15.86	0.04
`Minmax-entropy`	91.61	352.36	**8.33**	3.43	7.50	9.10	**11.51**	26.61	16.23	7.22
`EigenRatio`	43.95	1.48	27.77	0.02	9.01	0.03	N/A	N/A	N/A	N/A
`KOS`	51.95	9.98	**11.11**	0.01	39.75	0.03	42.93	0.31	31.84	0.13
`GhoshSVD`	43.03	11.62	27.77	0.01	49.12	0.03	N/A	N/A	N/A	N/A
`Majority Voting`	34.85	N/A	21.29	N/A	10.31	N/A	26.93	N/A	17.91	N/A

References

- Dawid, A. P. and Skene, A. M. *Maximum likelihood estimation of observer error-rates using the em algorithm.* Applied statistics, pp. 20–28, 1979.
- Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. *Spectral methods meet em: A provably optimal algorithm for crowdsourcing.* In Advances in Neural Information Processing Systems, pp. 1260–1268, 2014.
- Donoho, D. and Stodden, V. *When does non-negative matrix factorization give the correct decomposition into parts?* In Advances in Neural Information Processing Systems, pp. 1141-1148 2004.