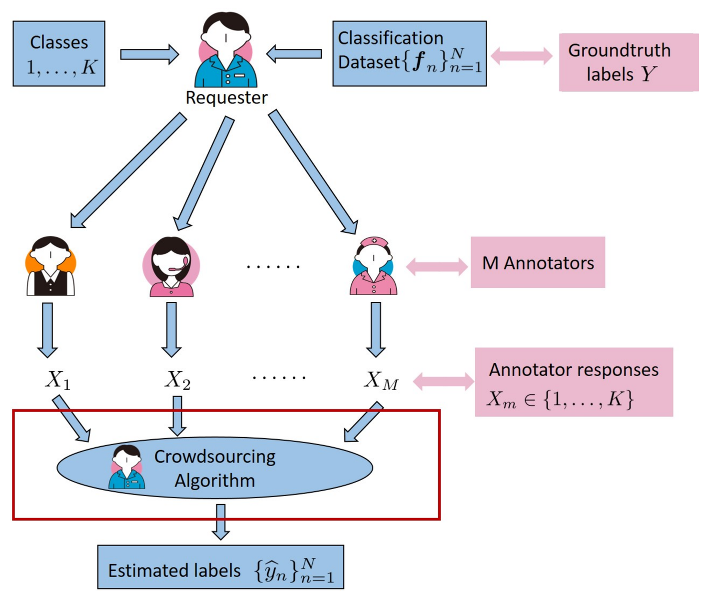
Crowdsourcing via Annotator Co-occurrence Imputation and Provable Symmetric Nonnegative Matrix Factorization

Shahana Ibrahim, Xiao Fu

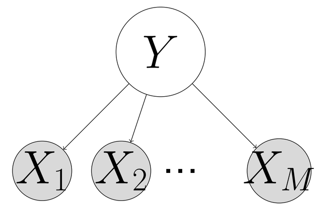
School of EECS, Oregon State University, Corvallis, OR, USA

Crowdsourcing

- **Crowdsourcing** techniques
  - ☐ employ a group of annotators to label the data items
  - ☐ integrate the acquired labels



Dawid-Skene Model



- Dawid and Skene [1979] formulated label integration as **model identification**
- Under Dawid & Skene (D&S) model,

$$\Pr(X_1 = k_1, \dots, X_M = k_M) = \sum_{k=1}^{K} \Pr(Y = k) \prod_{m=1}^{M} \Pr(X_m = k_m | Y = k)$$

- Define the **confusion matrix** $\boldsymbol{A}_m \in \mathbb{R}^{K\times K}$ for each annotator and the **prior probability vector** $\boldsymbol{\lambda} \in \mathbb{R}^K$ such that

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

- One can build a MAP estimator for y_n after identifying $\boldsymbol{A}_m$'s and $\boldsymbol{\lambda}$

Prior Approaches with Dawid-Skene Model

- Dawid-Skene (D&S) Model & EM Algorithm [Dawid and Skene, 1979]:
  - ☐ No model identifiability & algorithm tractability
- Bayesian Methods [Whitehill et al., 2009; Zhou et al., 2012]:
  - Extended D&S model considering "item difficulty" and "annotator ability"
  - No model identifiability
- Tensor Methods [Zhang et al., 2016; Traganitis et al., 2018]:
  - ☐ Using third-order co-occurrences of annotator responses, for e.g., $\Pr(X_m = k_m, X_\ell = k_\ell, X_j = k_j)$
  - ☐ Established model identifiability
  - ☐ High sample complexity due to third-order statistics
  - ☐ High computational cost from the tensor decomposition
- Coupled NMF (CNMF)-based Approach [Ibrahim et al., 2019]:
  - ☐ Using **pairwise co-occurrences** of responses: $\boxed{\boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top, \boldsymbol{D} = \mathrm{diag}(\boldsymbol{\lambda})}$

$$\underbrace{\Pr(X_m = k_m, X_j = k_j)}_{\boldsymbol{R}_{m,j}(k_m,k_j)} = \sum_{k=1}^{K} \underbrace{\Pr(Y = k)}_{\boldsymbol{\lambda}(k)} \underbrace{\Pr(X_m = k_m | Y = k)}_{\boldsymbol{A}_m(k_m,k)} \underbrace{\Pr(X_j = k_j | Y = k)}_{\boldsymbol{A}_j(k_j,k)}$$

  - ☐ less sample complexity compared to third-order statistics
  - If annotators m and j **co-label** some items, $\boldsymbol{R}_{m,j}$ can be estimated via sample averaging

CNMF Approach - A Deeper Look

- The CNMF criterion in [Ibrahim et al., 2019]:

$$\begin{aligned} &\text{find } \{\boldsymbol{A}_m\}_{m=1}^M, \boldsymbol{\lambda} \\ &\text{s.t. } \boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top,\ (m,j) \in \boldsymbol{\Omega}, \leftarrow \textit{observed set} \\ &\quad \boldsymbol{A}_m \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{A}_m = \boldsymbol{1}^\top,\ \boldsymbol{1}^\top \boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{aligned}$$

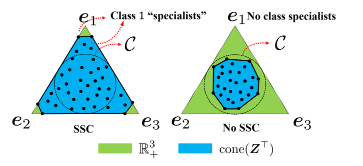
- Identifiability under the assumption that there exist two subsets of annotators $\mathcal{P}_1$ and $\mathcal{P}_2$, where $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$ and $\mathcal{P}_1 \cup \mathcal{P}_2 \subseteq [M]$,

$$\boldsymbol{H}^{(1)} := [\boldsymbol{A}_{m_1}^\top, \dots, \boldsymbol{A}_{m_{|\mathcal{P}_1|}}^\top]^\top, \quad \boldsymbol{H}^{(2)} := [\boldsymbol{A}_{j_1}^\top, \dots, \boldsymbol{A}_{j_{|\mathcal{P}_2|}}^\top]^\top,$$

s.t. $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ satisfy the *sufficiently scattered condition* (SSC)

Def. 1: (SSC) [Fu et al., 2015]

Any nonnegative matrix $\boldsymbol{Z} \in \mathbb{R}_+^{I\times K}$ satisfies the SSC if the conic hull of $\boldsymbol{Z}^\top$ (i.e., cone($\boldsymbol{Z}^\top$)) satisfies $\mathcal{C} \subseteq \mathrm{cone}\{\boldsymbol{Z}^\top\}$ where $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^K \mid \boldsymbol{x}^\top \boldsymbol{1} \geq \sqrt{K-1}\|\boldsymbol{x}\|_2\}$.



- A row of $\boldsymbol{H}^{(i)}$ (i.e., a row of $\boldsymbol{A}_m$) close to kth unit vector implies
  $\boldsymbol{A}_m(k,k) \approx 1$ and $\boldsymbol{A}_m(k,k_m) \approx 0, k_m \neq k$ **(class specialists)**,
  i.e., annotator m rarely confuses data from other classes with those from class k
- **Identifiability Challenge:**
  - Both $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ satisfy the SSC $\implies$ the disjoint $\mathcal{P}_1$ and $\mathcal{P}_2$ both contain "**class specialists**" for all K classes (somewhat restrictive condition)
- **Computational Challenges:**
  - The CNMF criterion in [Ibrahim et al., 2019] is handled using KL-divergence based model fitting problem with constraints (hardly scalable)
  - Unclear convergence guarantee even if there is no noise
  - Unclear identifiability guarantee when there is noise

Proposed Approach - SymNMF Framework

- Assume that all $\boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top$ are available for all $m, j \in [M]$

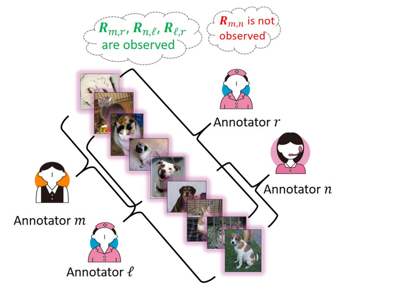
Symmetric Non-negative Matrix Factorization (SymNMF)

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{R}_{1,1} & \dots & \boldsymbol{R}_{1,M} \\ \vdots & \ddots & \vdots \\ \boldsymbol{R}_{M,1} & \dots & \boldsymbol{R}_{M,M} \end{bmatrix} = \underbrace{[\boldsymbol{A}_1^\top, \dots \boldsymbol{A}_M^\top]^\top \boldsymbol{D}^{1/2}}_{\boldsymbol{H}} \underbrace{\boldsymbol{D}^{1/2}[\boldsymbol{A}_1^\top, \dots \boldsymbol{A}_M^\top]}_{\boldsymbol{H}^\top}$$

- **If $\boldsymbol{H}$ satisfies SSC**, the SymNMF model is unique [Huang et al., 2014], i.e., **$\boldsymbol{A}_m$'s and $\boldsymbol{\lambda}$ can be identified upto common column permutations**
- SSC of $\boldsymbol{H}$ $\implies$ only one set of "**class specialists**" is needed
  - the CNMF framework needs two disjoint sets of annotators $\mathcal{P}_1$ and $\mathcal{P}_2$ both contain "**class specialists**" for all K classes
  - much easier to satisfy compared to the CNMF framework case
- The challenge in SymNMF framework is that many $\boldsymbol{R}_{m,j}$'s may be missing:
  - $\boldsymbol{R}_{m,m} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_m^\top, \forall m$ **do not have physical meaning** and thus cannot be observed
  - if annotators m, j **never co-labeled any items**, $\boldsymbol{R}_{m,j}$ is missing

Designated Annotators-based Imputation

- In crowdsourcing, some annotators may be *designated* to co-label items with other annotators.



1. $\boldsymbol{C} \longleftarrow [\boldsymbol{R}_{m,r}^\top, \boldsymbol{R}_{\ell,r}^\top]^\top$
2. $\boldsymbol{C} \xrightarrow{\text{thin SVD}} [\boldsymbol{U}_m^\top, \boldsymbol{U}_\ell^\top]^\top \boldsymbol{\Sigma}_{m,\ell,r} \boldsymbol{V}_r^\top$
3. $\boldsymbol{R}_{m,n} \longleftarrow \boldsymbol{U}_m \boldsymbol{U}_\ell^{-1} \boldsymbol{R}_{n,\ell}^\top$

The diagonal blocks $\boldsymbol{R}_{m,m}$'s can be estimated by observing $\boldsymbol{R}_{m,\ell}$, $\boldsymbol{R}_{m,r}$, and $\boldsymbol{R}_{\ell,r}$

Theorem 1: (Informal)

Assume that $\boldsymbol{R}_{m,r}$, $\boldsymbol{R}_{n,\ell}$ and $\boldsymbol{R}_{\ell,r}$ are estimated using at least S items and that $\kappa(\boldsymbol{A}_m) \leq \gamma$ and $\mathrm{rank}(\boldsymbol{A}_m) = \mathrm{rank}(\boldsymbol{D}) = K$ for all m. Suppose that S is above certain threshold. Then, any unobserved $\boldsymbol{R}_{m,n}$ can be estimated via (1)-(3), with probability of at least $1-\delta$ such that $\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} = O\left(K^2\gamma^3\sqrt{\log(1/\delta)/S}\right)$.

- **What if we do not have designated annotators?**

Robust Co-occurrence Imputation Criterion

$$\begin{aligned} \underset{\boldsymbol{U}_m, \boldsymbol{U}_j,\ \forall (m,j)\in\boldsymbol{\Omega}}{\text{minimize}} \quad & \sum_{(m,j)\in\boldsymbol{\Omega}} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m \boldsymbol{U}_j^\top\|_{\mathrm{F}} \\ \text{subject to } \quad & \|\boldsymbol{U}_m\|_{\mathrm{F}} \leq D_{(\text{certain upper bound})}, \ \forall m \end{aligned}$$

- block ℓ_2/ℓ_1-mixed norm based criterion
- $\widehat{\boldsymbol{R}}_{m,j}$'s are estimated using unequal no of samples
- the formulation is robust under such unbalanced estimates

Theorem 2: Stability under Finite Samples

Assume that $\widehat{\boldsymbol{R}}_{m,j}$'s are estimated with $S_{m,j}$ samples, $\forall\ (m,j) \in \boldsymbol{\Omega}$ and each $\widehat{\boldsymbol{R}}_{m,j}$ is observed with the same probability. Let $\{\boldsymbol{U}_m^*, \boldsymbol{U}_j^*\}$ be any optimal solution of the above. Then we have

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

with probability of at least $1 - 3\exp(-M)$, where $L = M(M-1)/2$ and $C > 0$.

Shifted ReLU Empowered SymNMF

Assuming that $\boldsymbol{X}$ is observed after co-occurrence imputation:

$$\boxed{\boldsymbol{X} = \boldsymbol{H}\boldsymbol{H}^\top \xrightarrow{\text{square root decomp.}} \boldsymbol{X} \rightarrow \boldsymbol{U}\boldsymbol{U}^\top \implies \boldsymbol{U} = \boldsymbol{H}\boldsymbol{Q}^\top, \boldsymbol{Q} \text{ is orthogonal}}$$

Estimation Criterion:

$$\begin{aligned} \underset{\boldsymbol{H},\boldsymbol{Q}}{\text{minimize}} \quad & \|\boldsymbol{H} - \boldsymbol{U}\boldsymbol{Q}\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \boldsymbol{H} \geq \boldsymbol{0},\ \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I} \end{aligned}$$

Proposed Algorithm:

$$\boldsymbol{H}_{(t+1)} \leftarrow \mathrm{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{U}\boldsymbol{Q}_{(t)}\right) \ \textbf{(Orth. proj. of each element of } \boldsymbol{U}\boldsymbol{Q}_{(t)} \textbf{ to } [\alpha_{(t)}, +\infty))$$

$$\left.\begin{aligned} &\boldsymbol{W}_{(t+1)}\boldsymbol{\Sigma}_{(t+1)}\boldsymbol{V}_{(t+1)}^\top \leftarrow \mathrm{svd}\left(\boldsymbol{H}_{(t+1)}^\top \boldsymbol{U}\right) \\ &\boldsymbol{Q}_{(t+1)} \leftarrow \boldsymbol{V}_{(t+1)}\boldsymbol{W}_{(t+1)}^\top \end{aligned}\right\} \ \textbf{(Procrustes projection)}$$

- reminiscent of the SymNMF algorithm proposed in [Huang et al., 2014]:
  - always uses $\alpha_{(t)} = 0$; convergence w/wo noise is unclear
- Convergence analysis for SymNMF algorithms is challenging
  - most existing SymNMF works showed only stationary point convergence [Huang et al., 2014; He et al., 2011]

Convergence of the Proposed SymNMF Algorithm

Theorem 3: (Informal)

Consider $\widehat{\boldsymbol{U}} = \boldsymbol{H}\boldsymbol{Q}^\top + \boldsymbol{N}$. Denote $\nu = \|\boldsymbol{N}\|_{\mathrm{F}}$, $\sigma = \|\boldsymbol{H}\|_{\mathrm{F}}$, $h_{(t)} = \|\boldsymbol{H}_{(t)} - \boldsymbol{H}\boldsymbol{\Pi}\|_{\mathrm{F}}^2$ and $q_{(t)} = \|\boldsymbol{Q}_{(t)} - \boldsymbol{Q}\boldsymbol{\Pi}\|_{\mathrm{F}}^2$, where $\boldsymbol{\Pi}$ is any permutation matrix. Under the assumptions that,
☐ $\boldsymbol{H}$ is full rank and sparse; the energy of range space of $\boldsymbol{H}$ is well spread over its rows;
☐ the noise term ν and the initial error $q_{(0)}$ are small enough;
there exists $\alpha_{(t)} = \alpha > 0$, $\eta > 0$ and $0 < \rho < 1$ such that with high probability,

$$q_{(t)} \leq \rho q_{(t-1)} + O\left(K\sigma^2\nu^2\right), \quad h_{(t)} \leq 2\eta\sigma^2 q_{(t-1)} + 2\nu^2 \leftarrow \textbf{linear convergence}$$

- **Shifted ReLU** operator is crucial for guaranteeing the convergence
- The rate parameter ρ is smaller (faster convergence) if $\boldsymbol{H}$ is sparser

Experiment Results

- **Experiments - UCI Data:**
  - Each annotator (MATLAB classifiers) is allowed to label an item with prob. $p_m \in (0,1]$; randomly choosing two annotators and letting them label with higher prob. (i.e., p_d)

Table: UCI Connect4 dataset ($N = 20,561$, $M = 10$, $K = 3$)

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
DesSymNMF	33.45	**32.18**	**31.42**	0.061
DesSymNMF-EM	33.94	**32.50**	**31.40**	0.128
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
Majority Voting	37.76	36.88	36.75	-

- **Experiments - Amazon Mechanical Turk (AMT) Data:**

Table: AMT datasets "RTE" and "TREC"

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
DesSymNMF	13.87	3.32	36.75	71.31
DesSymNMF-EM	**7.25**	3.43	**29.36**	72.13
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
Majority Voting	10.31	N/A	34.85	N/A

References

- A. P. Dawid and A. M. Skene. *Maximum likelihood estimation of observer error-rates using the EM algorithm.* Applied statistics, pp 20-28, 1979.
- S. Ibrahim, X.Fu, N. Kargas, and K. Huang. *Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms.* In Advances in NeurIPS, vol 32, pp 7847-7857, 2019.