Under-Counted Tensor Completion with Neural Side Information Learner

Shahana Ibrahim, Xiao Fu,
Rebecca Hutchinson, Eugene Seo

Oregon State University

Under-counted Data

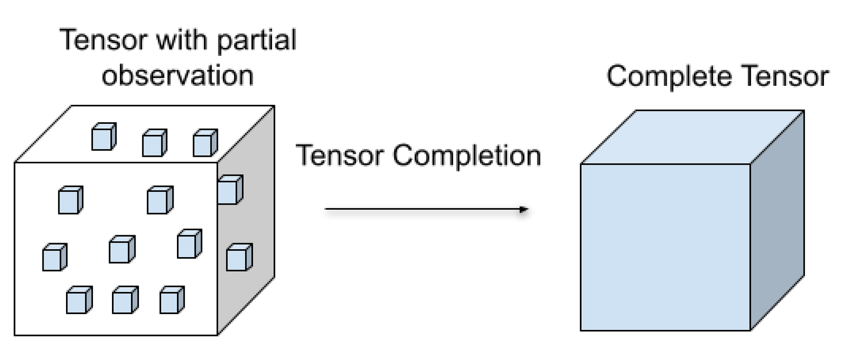
Plant-Pollinator Interactions

COVID-19 cases

- Under-counted data often arise in fields such as ecology and epidemiology
- An observer may very likely have just observed a small portion of the species' activities in some ecological datasets
- Actual number of infectious disease cases may be under-counted due to symptom-free patients or lack of testing in epidemiolocal datasets

Tensors Meet Under-counted Data

- Tensors are powerful tools for multi-aspect data analytics
- Tensor completion (TC) aims to recover a complete tensor from partial observations, often by leveraging its low rank structure
- Under-counted tensor completion (UC-TC) is less studied in literature



Key Components of the UC-TC Framework

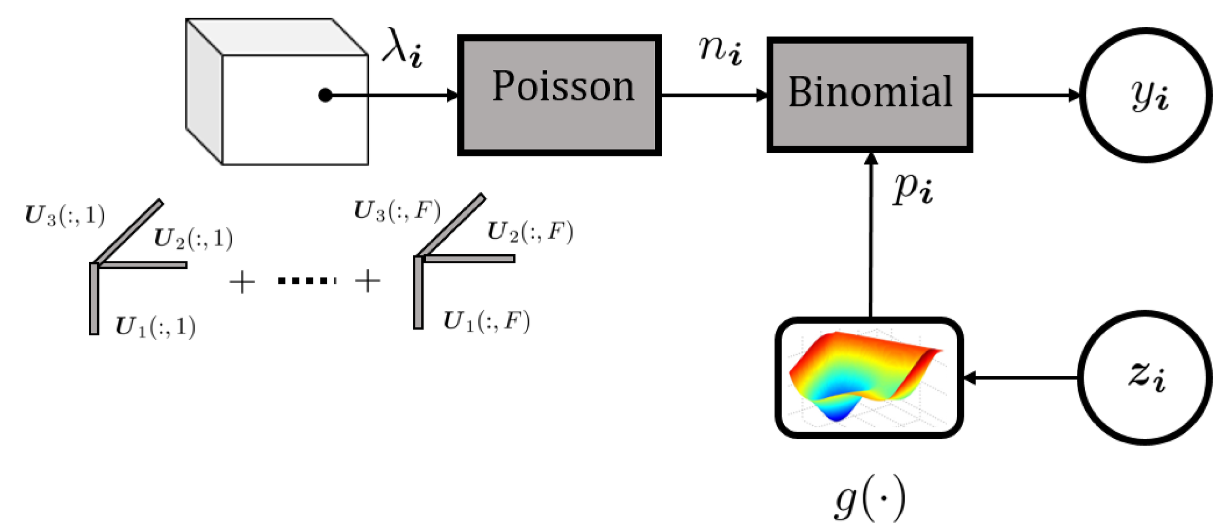
- y_i : Under-counted observations
- p_i : Probability of detection
- λ_i : Average true count
- $\boldsymbol{U}_1, \dots, \boldsymbol{U}_K$: Low rank tensor factors
- n_i : True counts
- $\boldsymbol{z}_i$: Side features, e.g., temperature, humidity, when observation is recorded
- $g(\cdot)$: Nonlinear function

Given a few under-counted observations y_i and a set of side features $\boldsymbol{z}_i$, can we recover the true counts and the detection probabilities for all tensor entries?

Proposed Approach

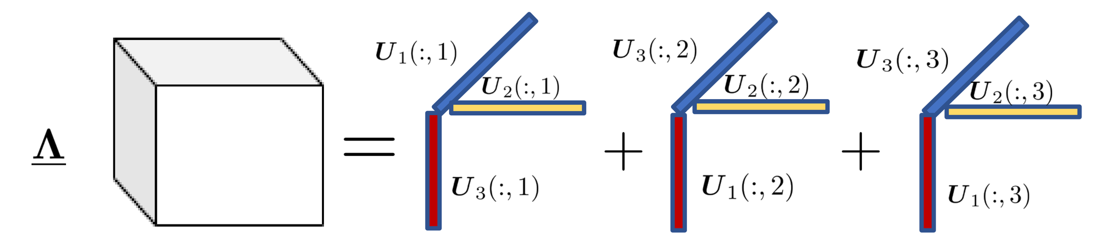
Proposed UC-TC Model

$$\lambda_i = \sum_{f=1}^{F} \prod_{k=1}^{K} \boldsymbol{U}_k(i_k, f),\ \boldsymbol{U}_k \in \mathbb{R}^{I_k \times F}, \boldsymbol{U}_k \geq \boldsymbol{0}, \forall k,$$
$$n_i \sim \text{Poisson}(\lambda_i),$$
$$p_i = g(\boldsymbol{z}_i),\ 0 \leq p_i \leq 1,$$
$$y_i \sim \text{Binomial}(n_i, p_i).$$



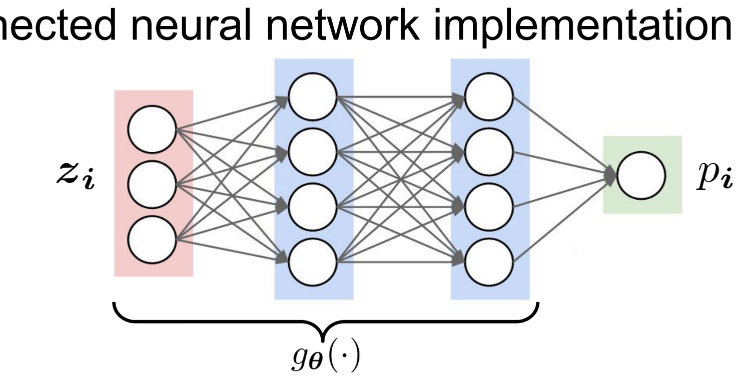
Proposed Uncle-TC Algorithm

- Maximum likelihood estimation (MLE)-based optimization
- Tensor factorization-based updates for $\boldsymbol{U}_1, \dots, \boldsymbol{U}_K$



Canonical Polyadic Decomposition – a low rank tensor factorization model

- Fully connected neural network implementation for $g(\cdot)$



Recoverability Analysis for UC-TC

- The first theory backed UC-TC method in literature
- Utilizes **diversity of observations & similarity of side features to show recoverability**

Key Analysis Results

- Estimation bound of average true counts: $|\lambda_i - \rho\widehat{\lambda}_i| \leq \eta_1,\ \forall i$
- Estimation bound of detection probabilities: $|p_i - \frac{1}{\rho}\widehat{p}_i| \leq \eta_2,\ \forall i$
- A global scaling ambiguity ρ between true count estimates and detection probability estimates

Related Work

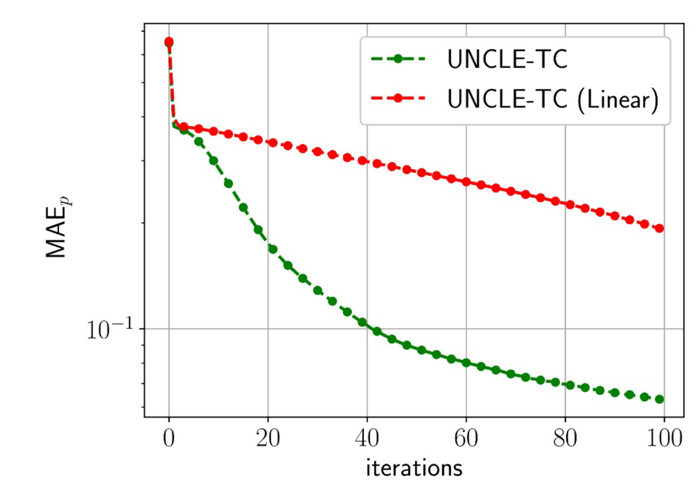
Under Counted Matrix Completion Model [Fu. et al., 2019]

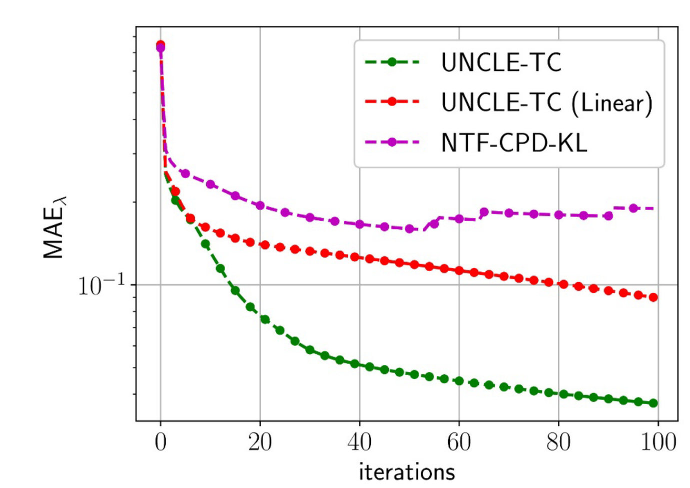
$$n_{i,j} \sim \text{Poisson}(\lambda_{i,j}),$$
$$\lambda_{i,j} = \boldsymbol{u}_i^\top \boldsymbol{v}_j,\ \boldsymbol{u}_i \geq \boldsymbol{0}, \boldsymbol{v}_j \geq \boldsymbol{0},$$
$$y_{i,j} \sim \text{Binomial}(n_{i,j}, p_{i,j}),\ p_{i,j} = \boldsymbol{z}_{i,j}^\top \boldsymbol{\theta}.$$

Limitations:

- Linear relation between side features and detection probabilities---do not capture complex nonlinear relationships
- Not supporting data having more than two aspects
- **No theoretical guarantees**

Synthetic Data Results





Real-Data Results

Results on Plant Pollinator Dataset

Method	rRMSE	AUROC	AUPRC
UncleTC	9.830	0.670	0.592
UncleTC (Linear)	10.862	0.657	0.542
HaLRTC	11.444	0.500	0.593
BPTF-CPD	10.852	0.665	0.501
NTF-CPD-KL	10.361	0.503	0.591
NTF-CPD-LS	11.252	0.597	0.454
NTF-Tucker-LS	11.196	0.621	0.456

Results on COVID-19 Dataset

Method	rRMSE	AUROC	AUPRC
UncleTC	1.834	0.596	0.921
UncleTC (Linear)	2.162	0.534	0.914
HaLRTC	4.399	0.501	0.911
BPTF-CPD	3.304	0.590	0.919
NTF-CPD-KL	3.399	0.564	0.918
NTF-CPD-LS	3.986	0.586	0.912
NTF-Tucker-LS	3.550	0.570	0.916

References

X. Fu, E. Seo, J. Clarke, and R. Hutchinson. Link prediction under imperfect detection: Collaborative filtering for ecological networks. IEEE Transactions on Knowledge and Data Engineering, 33(8):3117–3128, 2021.