Matrix Completion in the Unit Hypercube via Structured Matrix Factorization

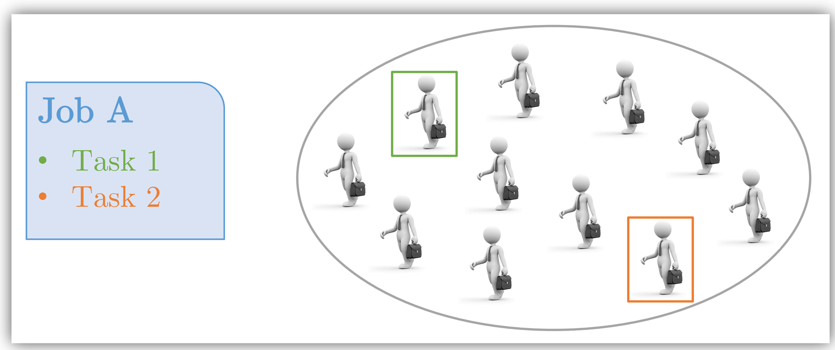
Emanuele Bugliarello[1], Swayambhoo Jain[2] and Vineeth Rakesh[2]
emanuele.bugliarello@nlp.c.titech.ac.jp,{Swayambhoo.Jain, Vineeth.Mohan}@technicolor.com

1 東京工業大学 Tokyo Institute of Technology
2 technicolor

IJCAI 2019

Motivation

In this study, we address a key challenge faced by several organizations: **how to select assets among available ones**. Specifically, given a *job A* consisting of two *tasks*, which employee should be assigned to each task?



Our key contributions are:

1. Transforming industrial setups into a real-valued matrix with entries in [0,1] representing employees' efficiencies
2. Proposing two novel structured matrix factorization (MF) models to solve the resulting matrix completion problem

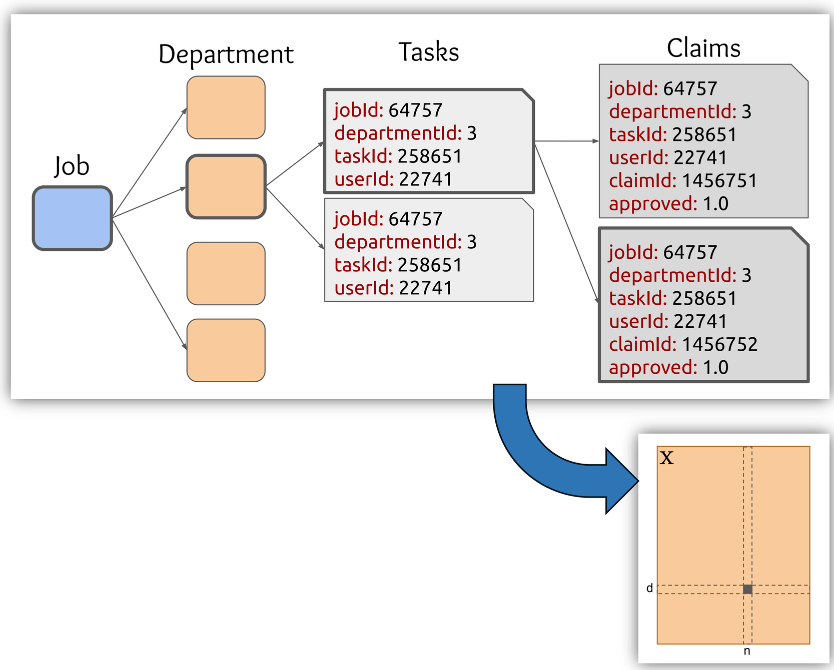
Experiments on three real-world datasets validate the effectiveness of our models.

Problem Formulation

- We consider a general framework where *employees* work in different *departments*
- Employees' partial results (called *claims*) are assessed by the department *manager*, who decides whether to approve them
- Employees' efficiencies can be measured as the ratio of accepted claims:

$$x_{dn} = \frac{\sum_{i=1}^{N_{dn}} a_{dn}^{(i)}}{N_{dn}} \in [0,1]$$

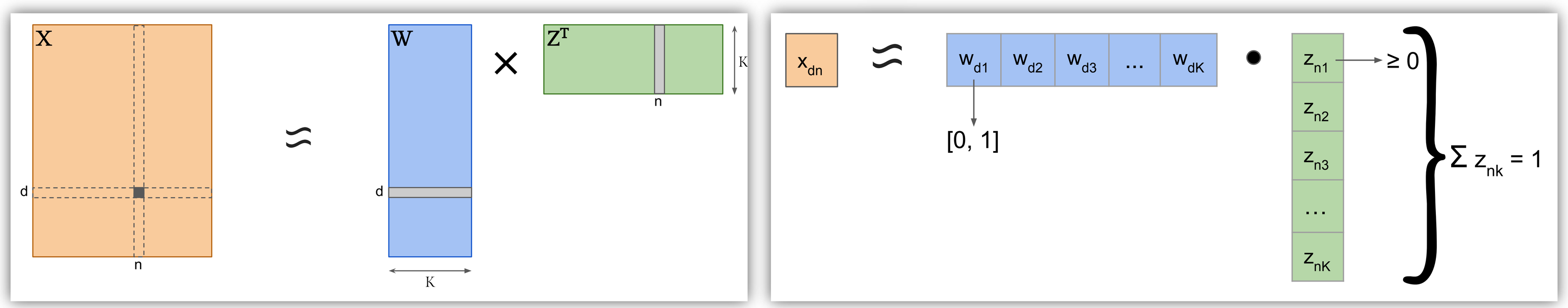
where
- N_{dn}: #claims by employee d in dept n
- $a_{dn}^{(i)}$: i-th claim of employee d in dept n was accepted (1) or rejected (0)



Our goal becomes to predict the missing entries of the efficiency matrix $\boldsymbol{X}$.

Proposed Models

Expertise Matrix Factorization



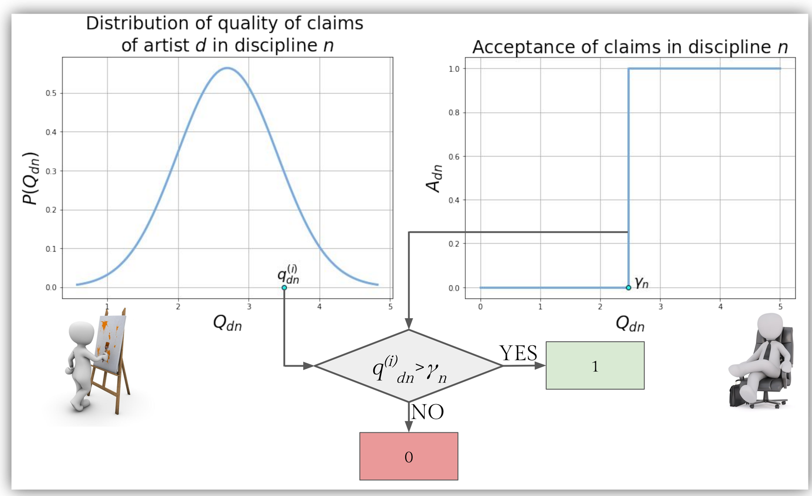
- *Expertise matrix factorization* (EMF) is a low-rank model $\boldsymbol{X} \approx \boldsymbol{W} \cdot \boldsymbol{Z}^T$ introducing structure on $\boldsymbol{W}$ and $\boldsymbol{Z}$ so that the entries of $\boldsymbol{W} \cdot \boldsymbol{Z}^T$ lie in [0, 1]
- We assume latent factors represent the *skills* required to work in a given dept
  - Ranging from 0 (no ability) to 1 (proficiency) for each employee
  - Being the distribution of skills required to complete tasks in each department
- That is, the efficiency of employee d in dept n is approximated by a weighted sum of the employee's skills and the importance of each skill in that department
- The resulting optimization problem is:

$$\min_{\boldsymbol{W}\geq 0, \boldsymbol{Z}\geq 0, \boldsymbol{\beta}\geq 0} \frac{1}{2|\Omega|} \sum_{(d,n)\in\Omega} [x_{dn} - (\beta_d + \boldsymbol{w}_d^T \cdot \boldsymbol{z}_n)]^2 + \frac{\lambda_u}{2}(\| \boldsymbol{W} \|_F^2 + \| \boldsymbol{\beta} \|_2^2) + \frac{\lambda_i}{2} \| \boldsymbol{Z} \|_F^2$$

$$\text{subject to} \quad \beta_d + w_{dk} \leq 1 \text{ and } \sum_{k=1}^{K} z_{nk} = 1, \quad \text{for } (d,k) \in [D]\times[K], n \in [N]$$

Survival Matrix Factorization

- *Survival Matrix Factorization* (SMF) is a probabilistic model for the process resulting in the efficiency values
- We assume each claim acceptance to be a Bernoulli random variable A_{dn} equal to 1 when $Q_{dn} > \gamma_n$:
  - Q_{dn}: Quality of claims submitted by employee d to department n
  - γ_n: Quality threshold of the manager of department n



- That is:

$$x_{dn} \approx \mathbb{P}[Q_{dn} > \gamma_n] = S_{Q_{dn}}(\gamma_n) = \int_{\gamma_n}^{+\infty} f_{Q_{dn}}(u)\, du,$$

where $S_{Q_{dn}}(\gamma_n)$ is the survival function of Q_{dn} at γ_n

- Assuming (i) employees' quality probability distributions to be normal and (ii) their means to be low-rank $\mu_{dn} \approx \boldsymbol{w}_d^T \cdot \boldsymbol{z}_n$, we solve the usual MF problem with:

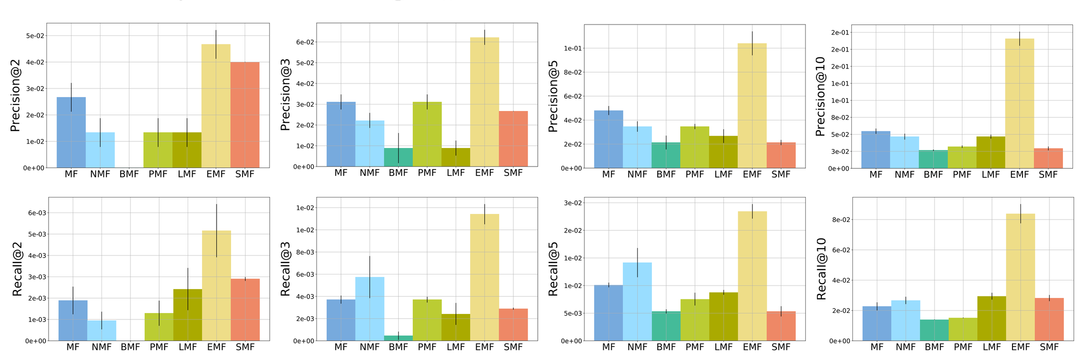
$$\hat{x}_{dn} = \int_{\gamma_n}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(u - \boldsymbol{w}_d^T \cdot \boldsymbol{z}_n)^2}{2\sigma^2}\right] du$$

Experimental Results

We evaluate the performance of our models in terms of Precision@N and Recall@N, for N ∈ {2, 3, 5, 10}. We compare our models against popular MF techniques, including Nonnegative matrix factorization (NMF), Bounded matrix factorization (BMF), Probabilistic matrix factorization (PMF) and Logistic matrix factorization (LMF).

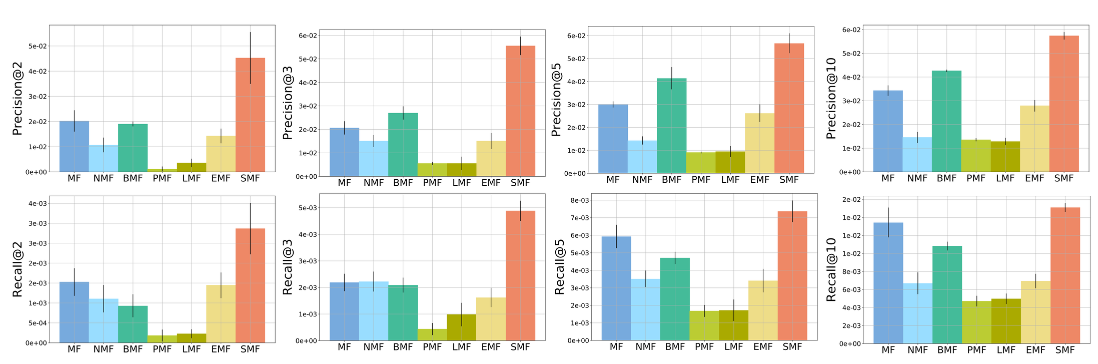
Movie Production

- Goal: Recommend artists to departments according to their efficiencies in rendering visual effects
- Data matrix $\boldsymbol{X}$:
  - 312 × 25
  - 86.85% sparse
  - very few ratings per user



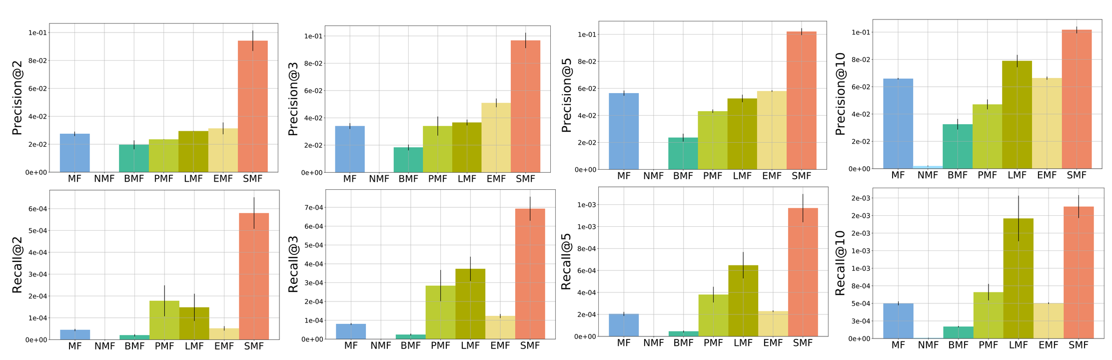
Over-The-Top

- Goal: Recommend apps to users in an Over-The-Top service according to their watching rates
- Data matrix $\boldsymbol{X}$:
  - 934 × 140
  - 99.91% sparse



Click-Through Rate

- Goal: Recommend website categories to an ad agency in order to maximize their click-through rate
- Data matrix $\boldsymbol{X}$:
  - 15647 × 85
  - 79.99% sparse
  - Wrangled from Outbrain Click Prediction



Conclusion

It is beneficial to explicitly model entries in the unit interval.

- EMF outperforms every method by at least 3× in the Movie Production dataset
  - The hypotheses of EMF closely match this framework, but not the others
  - EMF outperforms PMF in a dataset with a few ratings per user (max 7)
- SMF is a general model, outperforming every method in the other datasets
  - It underperforms in the Movie Production data due to very few available entries, not because of its sparsity (Over-The-Top data is much sparser)

References

- Experimental results: https://github.com/e-bug/unit-mf
- **BMF:** Ramakrishnan Kannan, Mariya Ishteva, and Haesun Park. Bounded matrix factorization for recommender system. *Knowledge and information systems*, 39(3):491–511, 2014.
- **PMF:** Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.
- **LMF:** Christopher C Johnson. Logistic matrix factorization for implicit feedback data. 2014.
- Outbrain Click Prediction: https://www.kaggle.com/c/outbrain-click-prediction