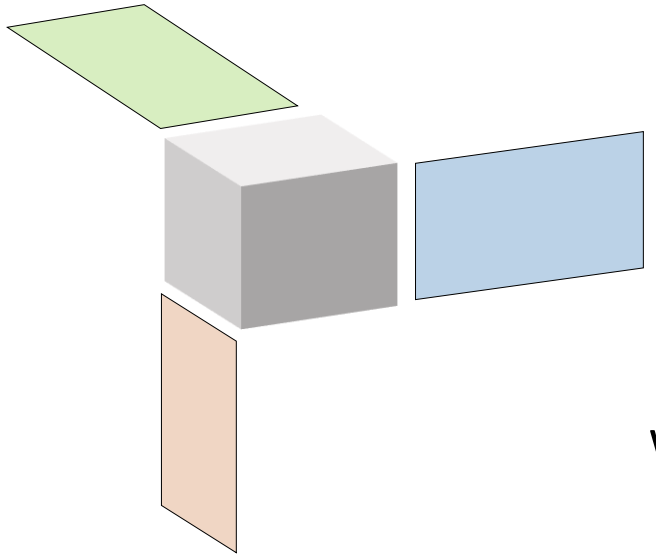
Max Planck Institute

Skoltech

Skolkovo Institute of Science and Technology

Generalization of low rank approximation approach for hybrid recommender systems



Evgeny Frolov

evgeny.frolov@skoltech.ru

Workshop on Low-rank Optimization and Applications, Leipzig
April 4, 2019

What is a recommender system?

Examples:

- Amazon
- Netflix
- Pandora
- Spotify
- etc.

Many different areas: e-commerce, news, tourism, entertainment, education…

Goal: predict user preferences given some prior information on user behavior.

Recommender systems

Engine

content-based

- user features
- item characteristics

collaborative filtering

- neighborhood-based approach
- model-based approach

hybrid systems

- broadly: ensembles and integration
- narrowly: content + collaborative data

Outcome

“utility” score

ranked list of items

Focus of my research

Collaborative filtering in real-world applications

Uses **collective information** about human behavior in order to predict **individual interests**.

This requires the ability to operate with **m(b)illions of users and items** and manage highly dynamic **online environments**.

Low rank matrix- and tensor-based models are especially suitable for this task and are **widely used in industry**.

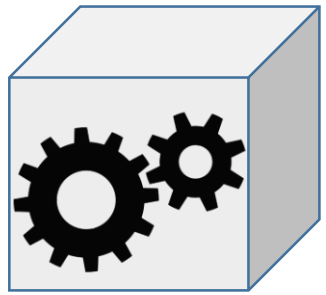
General workflow for collaborative filtering

Goal: predict user preferences based on prior user feedback and collective user behavior.

gather collective data

build model

f_U: $User \times Item \rightarrow Relevance\ Score$



f_U - utility function

generate recommendations

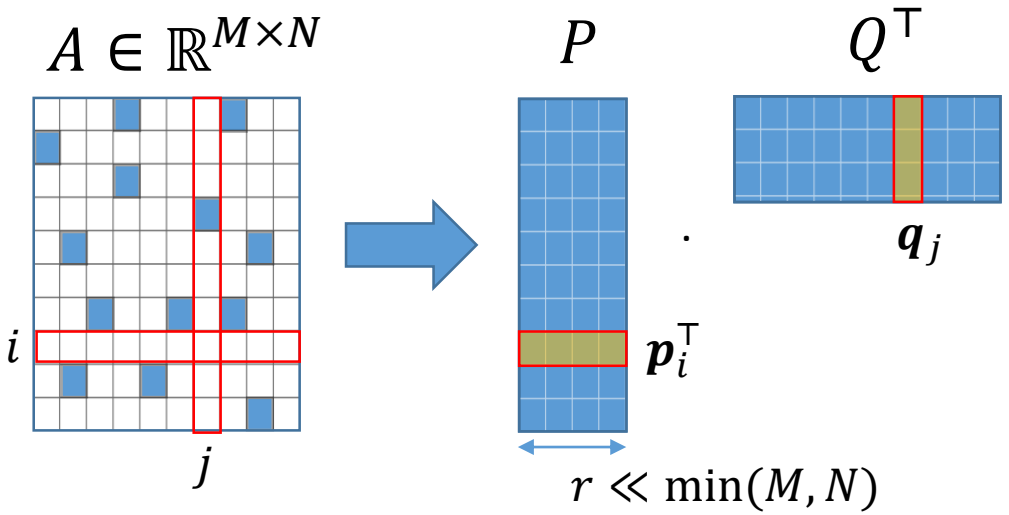
unknown user

A general view on matrix factorization

As optimization problem with some *loss function* $\mathcal{L}$:

$$\mathcal{L}(\text{historical data}, \text{model predictions}) \rightarrow \min$$

2D case: looking for solution in the form of a matrix factorization.



Some (oversimplified) intuition: latent features ↔ genres.

$f_U \rightarrow$ $r_{ij} = \boldsymbol{p}_i^\top \boldsymbol{q}_j$ predicted relevance of item j for user i

Generating top-n recommendations:

$$\text{rec}(i, n) = \arg\max_j^n r_{ij}$$

Dealing with incompleteness

$$\|W * (A - PQ^\top)\|_F^2 \to \min$$

Hadamard product

W masks unknowns: $w_{ij} = \begin{cases} 1, & a_{ij} \text{ is known,} \\ 0, & \text{otherwise.} \end{cases}$

elementwise form: $$\mathcal{J}(\mathrm{A}, \Theta) = \frac{1}{2} \sum_{i,j \in S} \left(a_{ij} - \boldsymbol{p}_i^T \boldsymbol{q}_j\right)^2 \qquad S = \{(i,j):\ w_{ij} \neq 0\}$$

full objective: $$\mathcal{L}(\Theta) = \mathcal{J}(\mathrm{A}, \Theta) + \Omega(\Theta)$$

$$\Theta = \{P, Q\}$$

additional constraints on factors

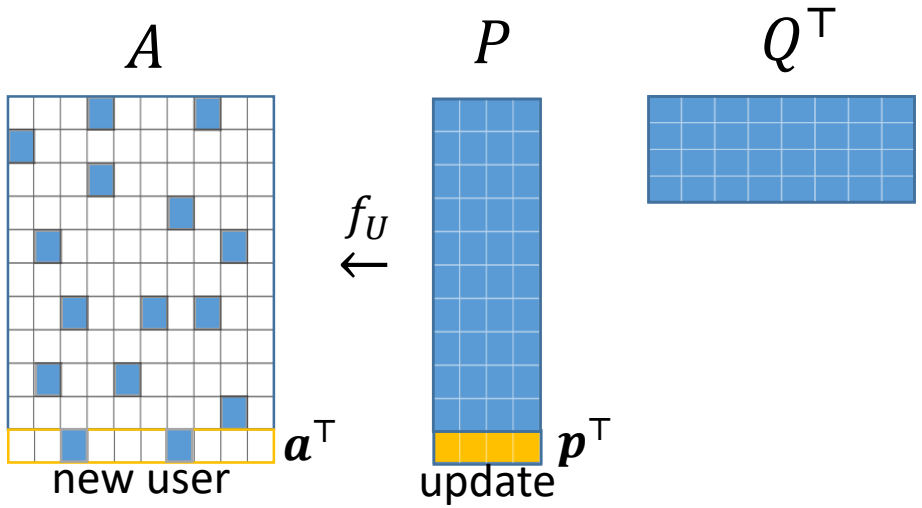
Typical optimization algorithms:

- stochastic gradient descent (SGD)
- alternating least squares (ALS)

Support for online settings

Task: provide instant recommendations to new or unrecognized users, assuming that at least a few interactions have occurred.

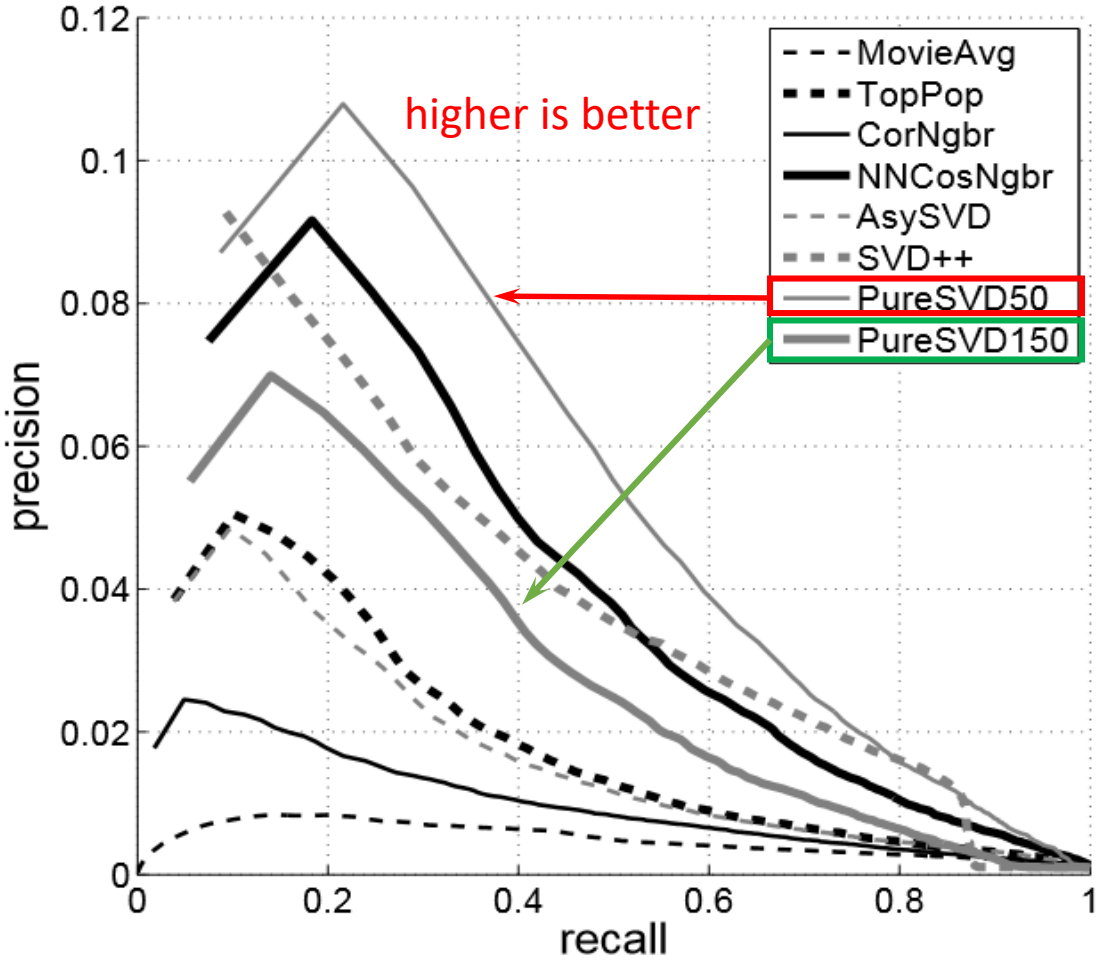
- Expected response time: <300ms.
- Common technique: *folding-in*.



$$\|\boldsymbol{a}^\top - Q^\top \boldsymbol{p}\|_2^2 \rightarrow \min$$

Matrix factorization in practice

Recommendations' quality of the most popular collaborative filtering techniques on the Netflix data.



Simple **PureSVD** model [Cremonesi/Koren/Turrin 2010]:

$$\|A_0 - R\|_F^2 \rightarrow \min,$$
$$\text{s.t.}\,\text{rank}(R) = r$$

unknowns are replaced with zeros in A_0.

vector of predicted item scores

$$\boldsymbol{p} = VV^{\top}\boldsymbol{a}_0$$

allows for real-time recommendations

PureSVD

Benefits:

- ✓simple tuning via rank truncation
- ✓supports dynamic online settings
- ✓stable, deterministic output
- ✓highly optimized implementations
- ✓scalability (randomized methods)

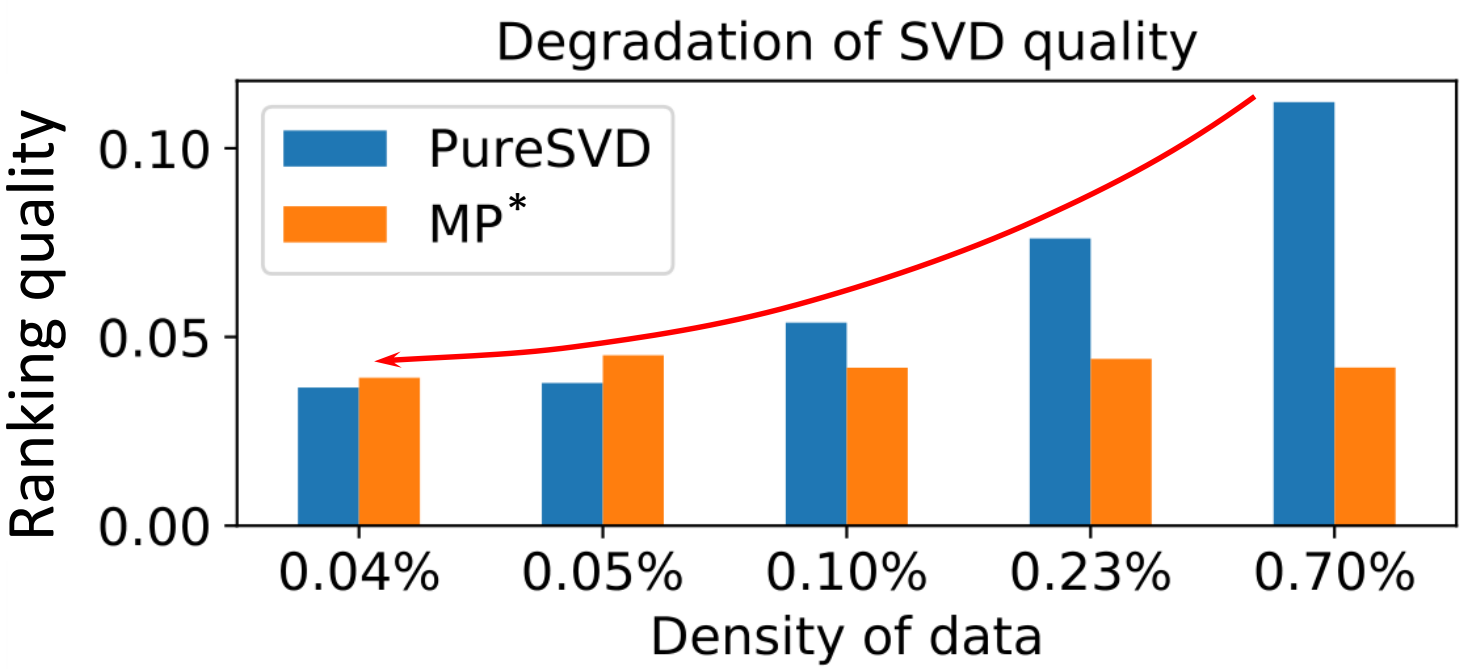
Issues:

- ❑ sensitive to the lack of observations
- ❑ no support for additional data:
  - user and item features – **content data**
  - situational information – **context data**

Remarks:

- we are solving a surrogate problem (not a ranking problem)
- not (strictly) a matrix completion

Data sparsity issue



*MP is a non-personalized popularity-based model

Content vs Context

What is the appropriate representation?

- **Content** is a *permanent* characterization of users or items;
  imposes structure on the lower dimensional latent space.

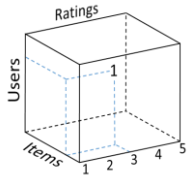
- **Context** characterizes *transient* relations between users and items;
  expands interaction space.

The main task is

to develop efficient low rank approximation model that:

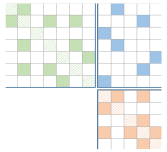
- inherits the key benefits of the PureSVD approach,
- is less susceptible to data sparsity,
- supports additional sources of information.

Roadmap



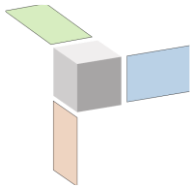
New higher order model

- Introduces key concepts of higher order approach to model context data.
- Based on the Tucker decomposition.
- Generalizable to any type of context.



New SVD-based hybrid model

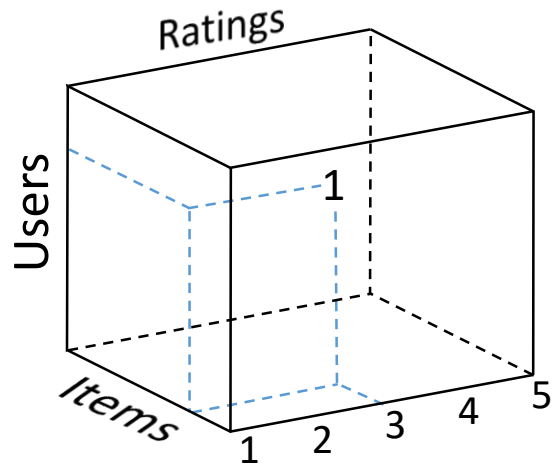
- Uses generalized SVD formulation to incorporate side information about users and items.
- Enjoys the benefits of the standard approach.
- More efficient learning over scarce interaction data.



Combined **hybrid tensor-based model**

- Combines the previous two methods.
- Addresses weak points of its predecessors.

1. Higher-order preference model

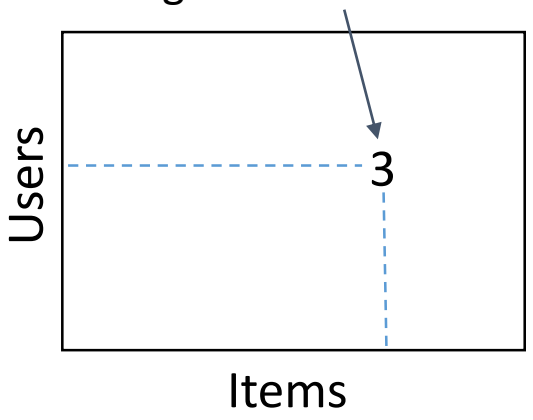


Restating the problem

Standard model

$User \times Item \rightarrow Rating$

ratings as **cardinal** values



$||A - X||_F^2 \rightarrow \min$

Technique: **Matrix factorization**

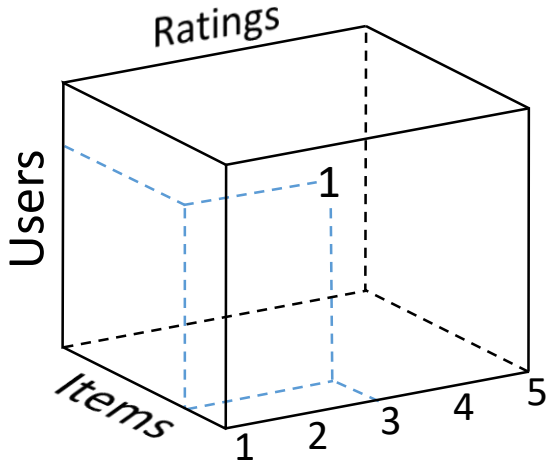
The model: $X \approx U\Sigma V^T$ (SVD)

Folding-in prediction vector: $\boldsymbol{p} = VV^T\boldsymbol{a}$

vector of known user preferences

Collaborative Full Feedback model – CoFFee (proposed approach)

$User \times Item \times Rating \rightarrow Relevance\ Score$



$||\mathcal{A} - \mathcal{X}||_F^2 \rightarrow \min$

Technique: **Tensor Factorization**

The model: $\mathcal{X} \approx \mathcal{G} \times_1 U \times_2 V \times_3 W$ (*Tucker Decomposition*)

Folding-in prediction matrix: $P = VV^T AWW^T$

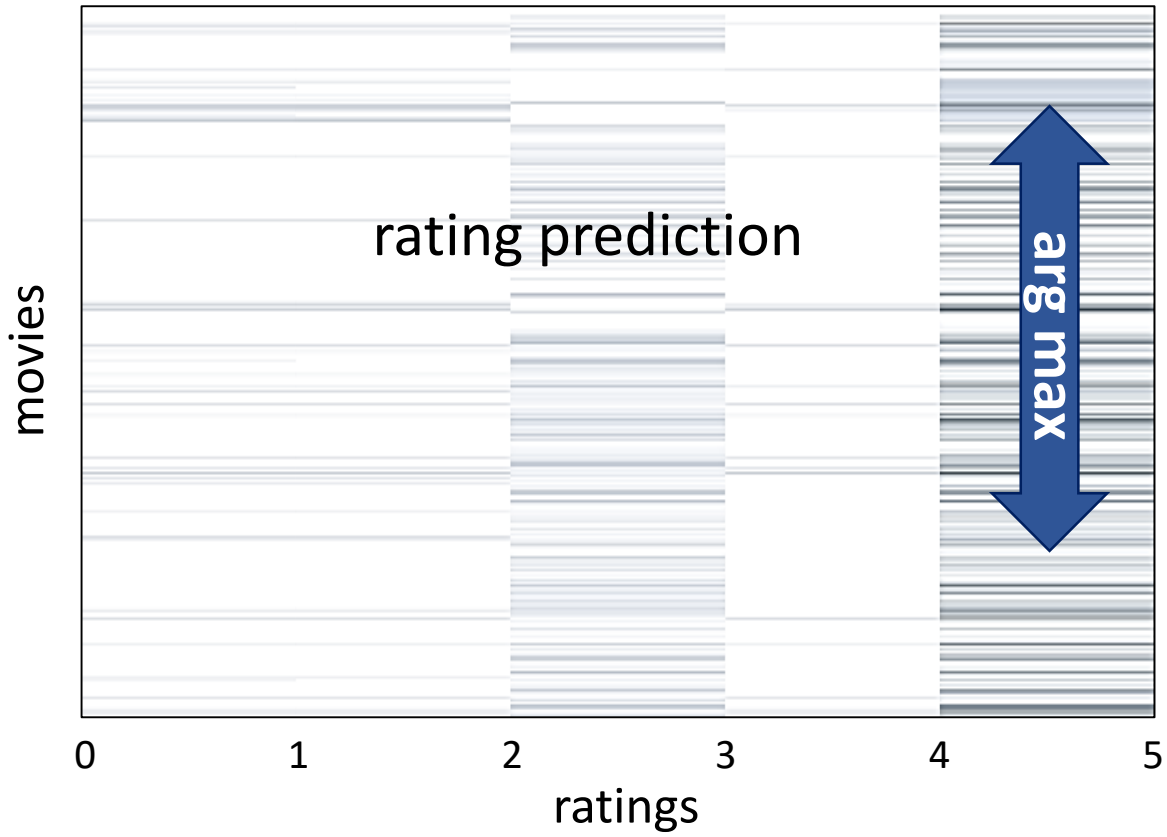
matrix of known user preferences

“Shades” of ratings

$$P = VV^T AWW^T$$

matrix of known user preferences

More dense colors correspond to higher relevance score.



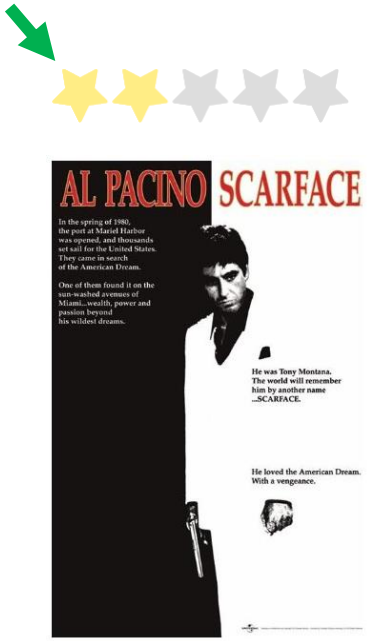
ranking task

Granular view of user preferences, concerning **all possible ratings**.

Model is **equally sensitive** to any kind of feedback.

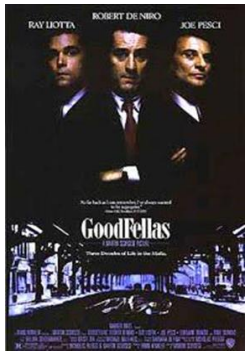
Warm start scenario

User feedback is negative!
Probably the user doesn't like criminal movies.



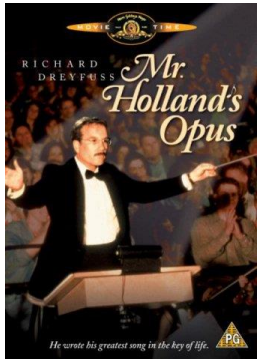
traditional methods

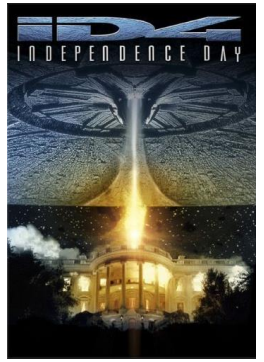
standard recommendations are insensitive to negative feedback



proposed method

our model predicts "opposite" preferences





Tensor approach uses the same amount of information, yet produces more expressive model.

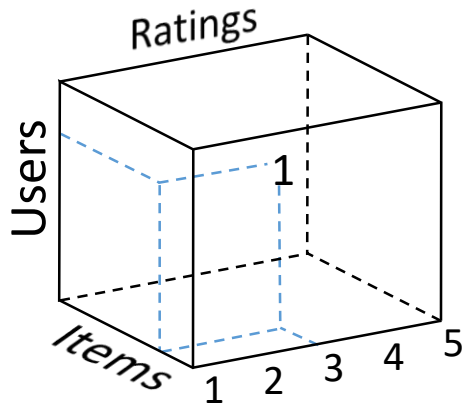
CoFFee – summary

✓Improves recommendations on *sufficiently dense* data.

✓Natural framework for including *contextual information*.

✓Supports quick online recommendations.

✓Offers simple rank tuning procedure (tensor rounding).

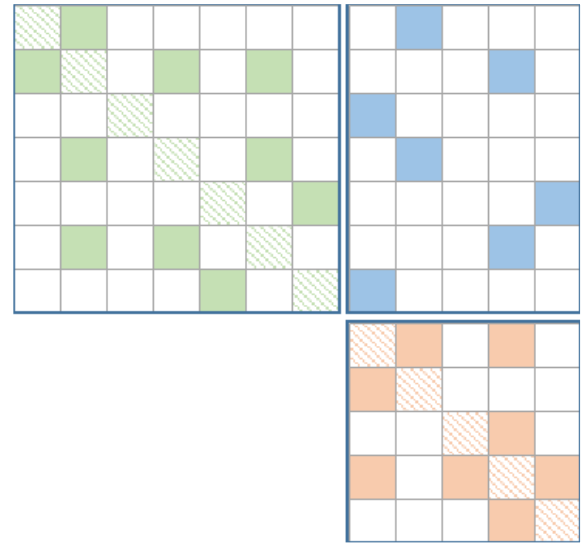


Shortcomings:

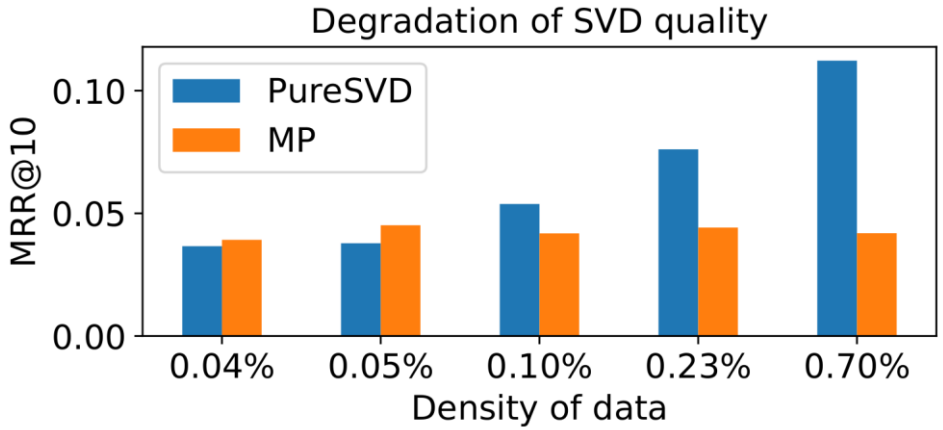
❑ Does not work for *content information* (e.g., movie genre).

❑ Suffers from extreme sparsity, *which it amplifies*.

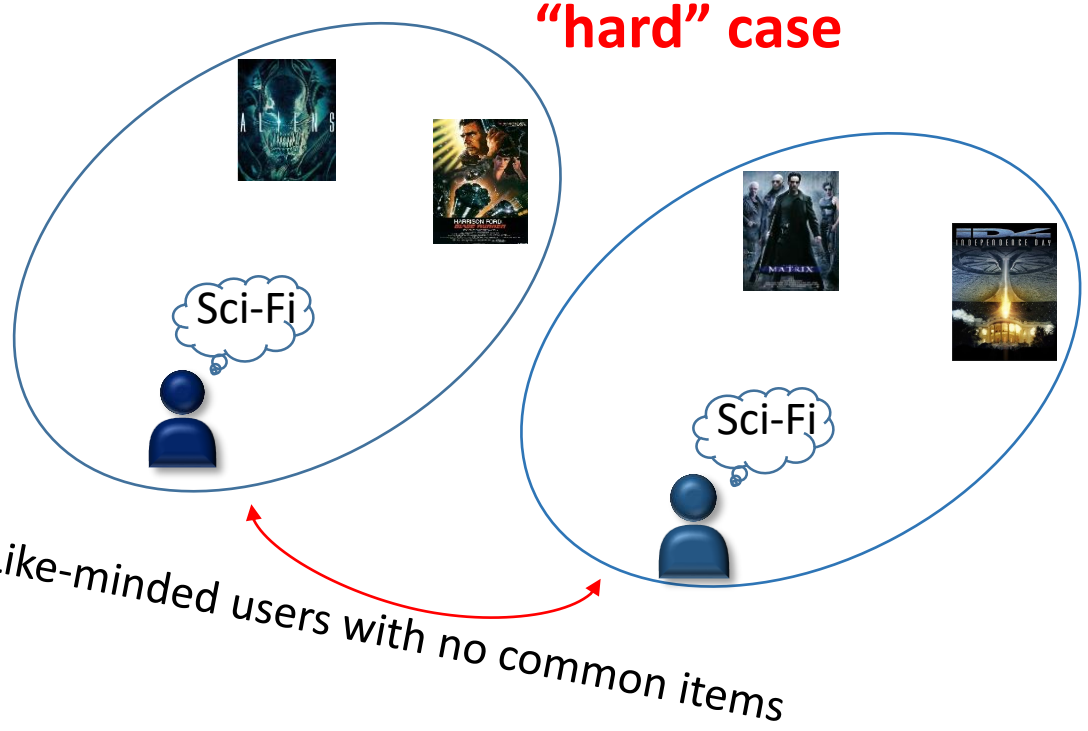
2. Hybrid SVD-based model



The problem of scarce interactions



VS.



Hybrid approach to include content

- Coupled (collective) factorization [Singh/Gordon 2008, Acar et al 2011],
- Local Collective Embeddidngs [Saveski/Mantrach 2014],
- Factorization Machines (polynomial expansion model) [Rendle 2010],
- Many others…

Example: coupled factorization.

$$\mathcal{L}(\Theta) = \|A - PQ^{\top}\|^2 + \|B - PU^{\top}\|^2 + \|C - QV^{\top}\|^2$$

$$\Theta = \{P, Q, U, V\}$$

matrix of user attributes

matrix of item features

- ❑ More flexible formulation at the cost of some computational benefits.
- ❑ Often leads to the growth of latent space.

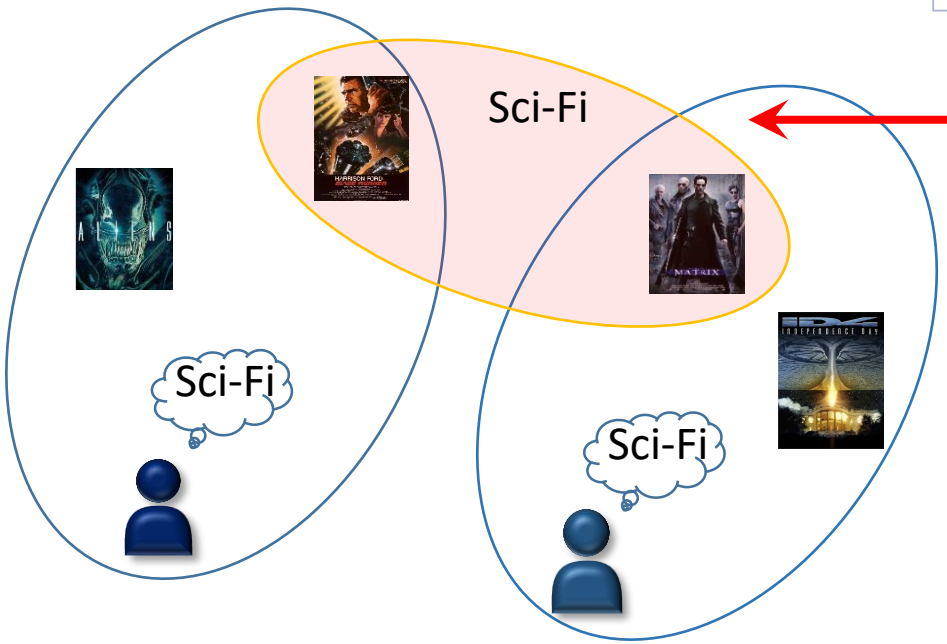
New approach - HybridSVD

PureSVD can be viewed as an *eigenproblem for the scaled cosine similarity matrix:*

$$AA^T = U\Sigma^2 U^T \quad \leftrightarrow \quad \mathrm{sim}(i,j) \sim a_i^T a_j \qquad a_i \text{ is an } i\text{-th row of a rating matrix } A$$

Key idea: replace scalar products with a bilinear form.

$$\mathrm{sim}(i,j) \sim a_i^T S\, a_j$$



"Virtual" connection based on item features encoded in matrix S.

Similarity matrix S

1			
	1	0.5	
	0.5	1	
			1

HybridSVD solution

From standard SVD of $A = U\Sigma V^{\mathsf{T}}$:

$$\begin{cases} AA^{\mathsf{T}} = U\Sigma^2 U^{\mathsf{T}} \\ A^{\mathsf{T}}A = V\Sigma^2 V^{\mathsf{T}} \end{cases} \Rightarrow \begin{cases} ASA^{\mathsf{T}} = U\Sigma^2 U^{\mathsf{T}} \\ A^{\mathsf{T}}KA = V\Sigma^2 V^{\mathsf{T}} \end{cases}$$

Solved via SVD of an auxiliary matrix: $\hat{A} \equiv L_K^{\mathsf{T}} A L_S = \hat{U}\Sigma\hat{V}^T,$

where $L_K L_K^T = K,\ L_S L_S^T = S.$

Connection to original latent space: $L_K^{-\mathsf{T}}\hat{U} = U,\ L_S^{-\mathsf{T}}\hat{V} = V$

Orthogonality property: $U^{\mathsf{T}}KU = I,\ \ V^{\mathsf{T}}SV = I.$

"Hybrid" folding-in: $\boldsymbol{p} = L_S^{-T}\hat{V}\hat{V}^T L_S^T \boldsymbol{a}.$

Remark on connection to probability theory

HybridSVD as an optimization problem:

$$\left\| L_K^{\top}(A - X)L_S \right\|_F^2 \to \min$$

or

$$\operatorname{tr}[K(A - X)S(A - X)^{\top}] \to \min.$$

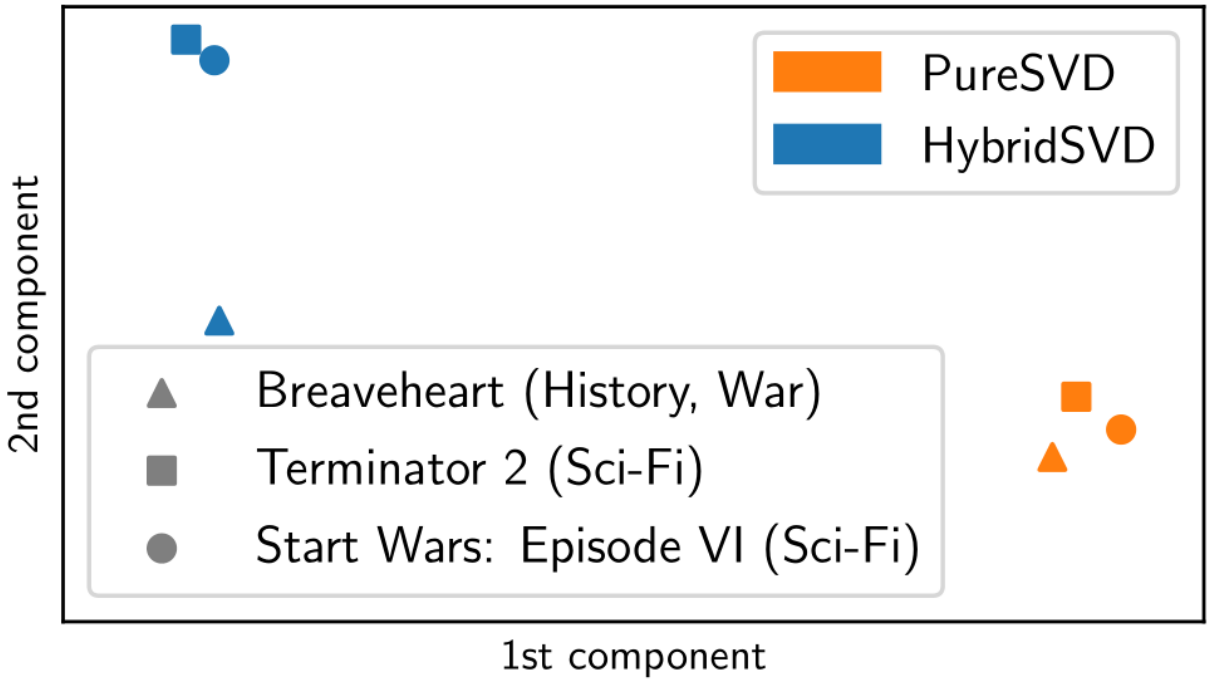
Corresponds to MLE for matrix-variate normal distribution (MVN).

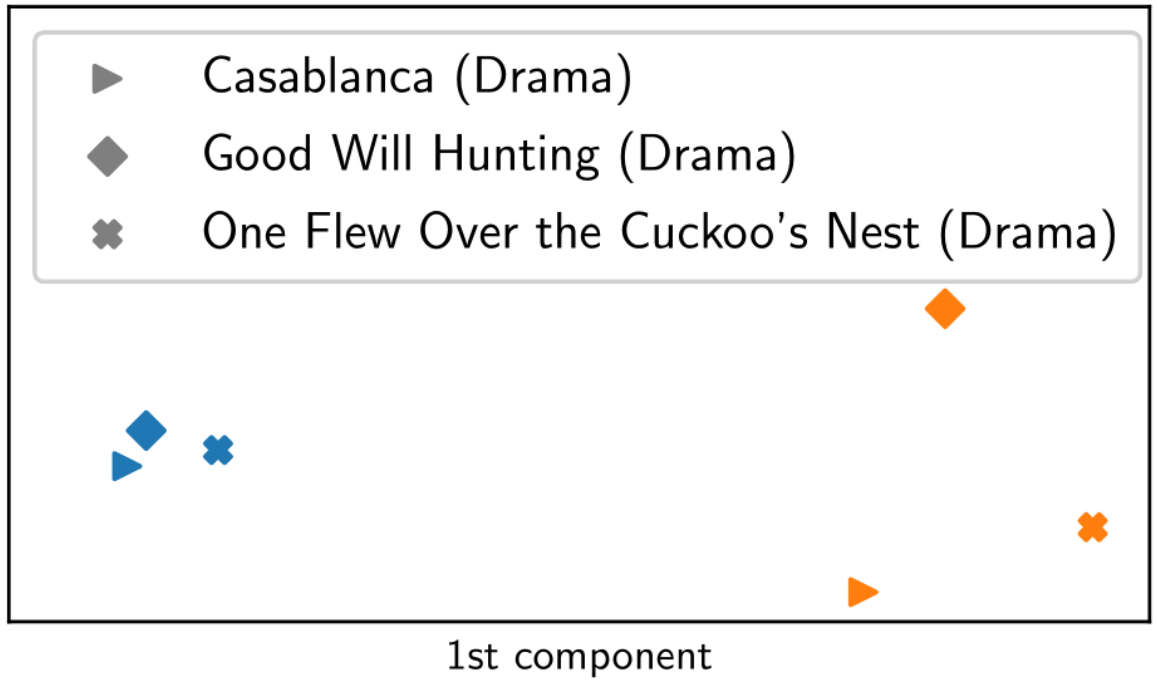
The density function for a random matrix X following the MVN distribution $\mathcal{MN}_{a\times b}(M, G, H)$ is

$$p(X) = \frac{\exp(-\frac{1}{2}\operatorname{tr}[G^{-1}(X - M)^{\top}H^{-1}(X - M)])}{(2\pi)^{ab/2}|G|^{a/2}|H|^{b/2}}$$

HybridSVD with movie genres

Construct similarity matrix S based on movie genres, set $K = I$.

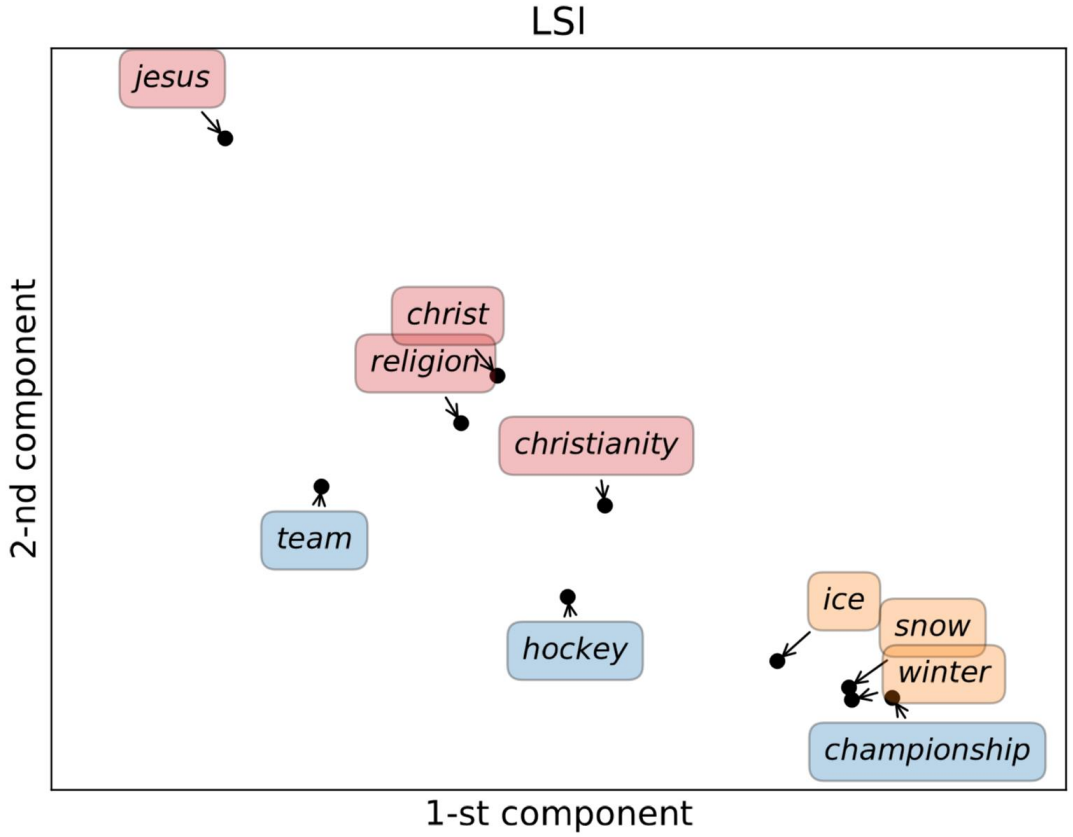


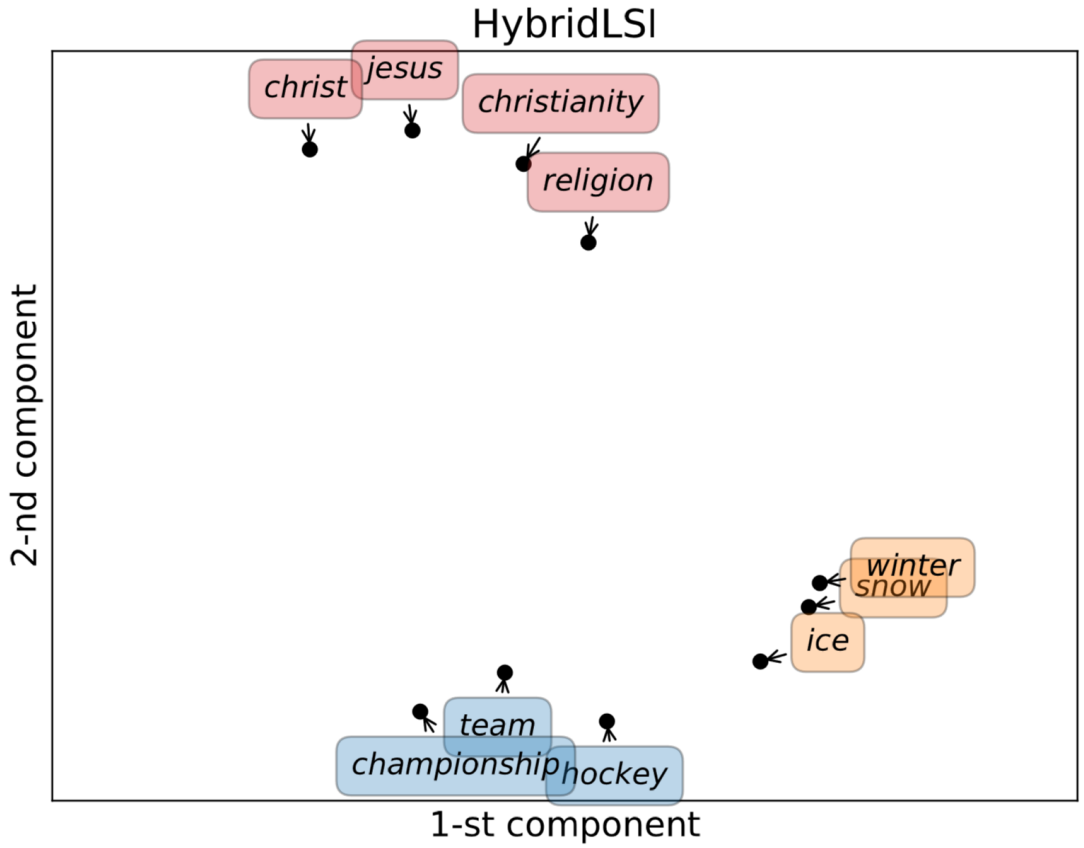


HybridSVD in document analysis

Use general semantic similarity of words based on a global model, e.g. word2vec.

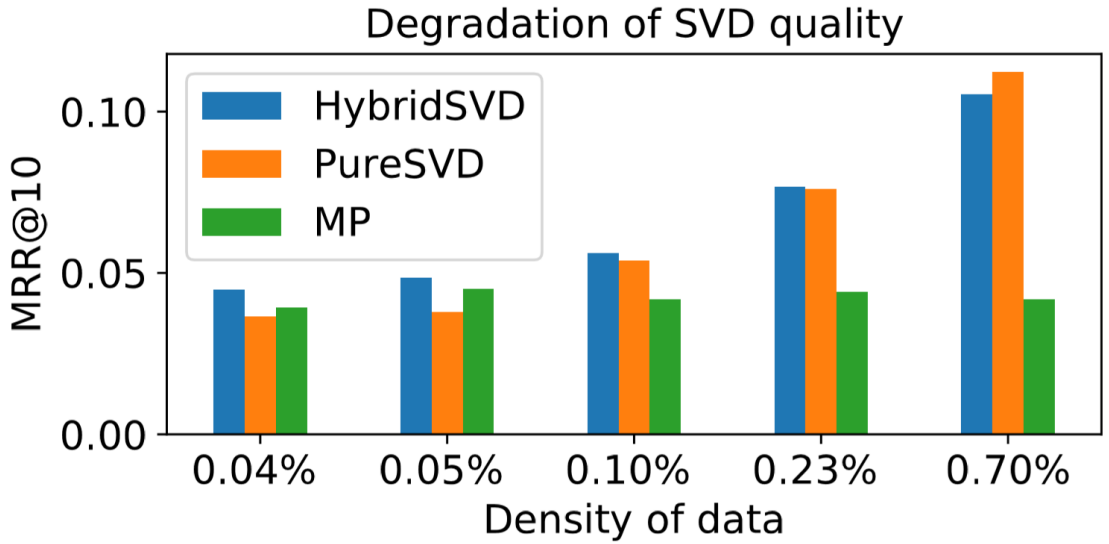
20 NewsGroups dataset

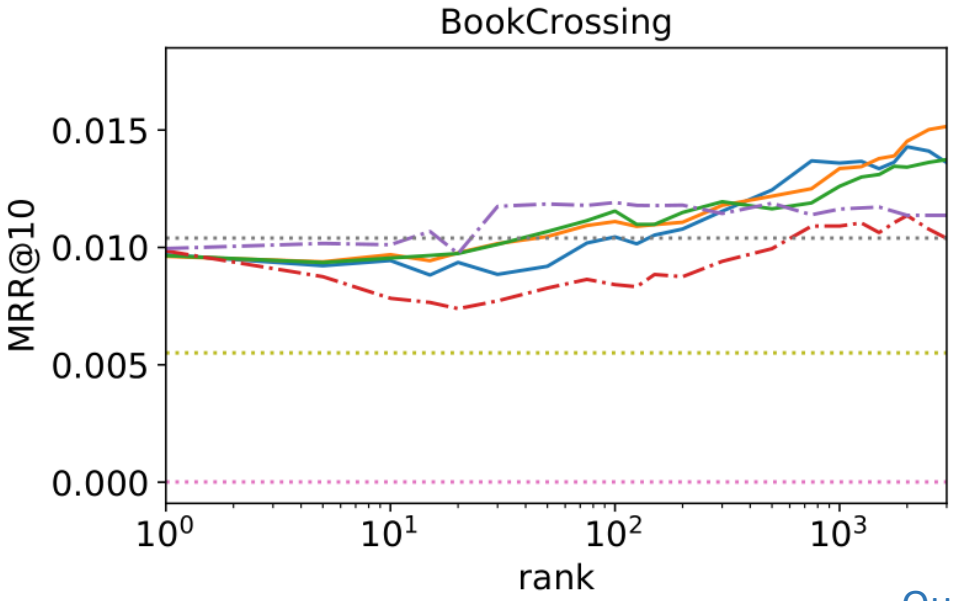


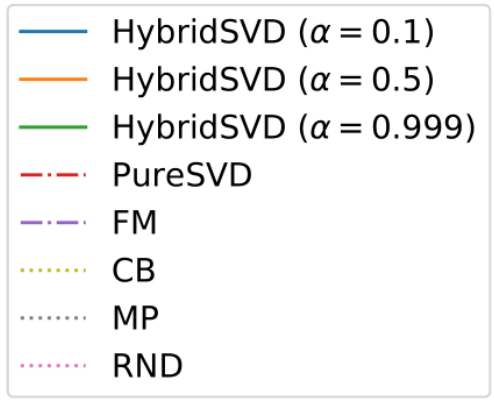


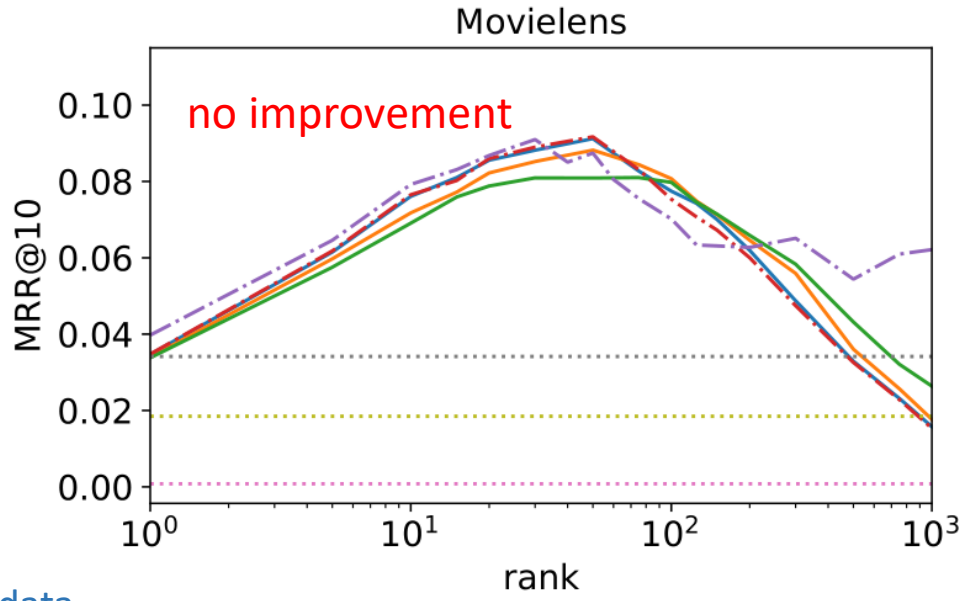
Images by Ilya Kolomeitsev, M.Sc., Skoltech

Intermediary results



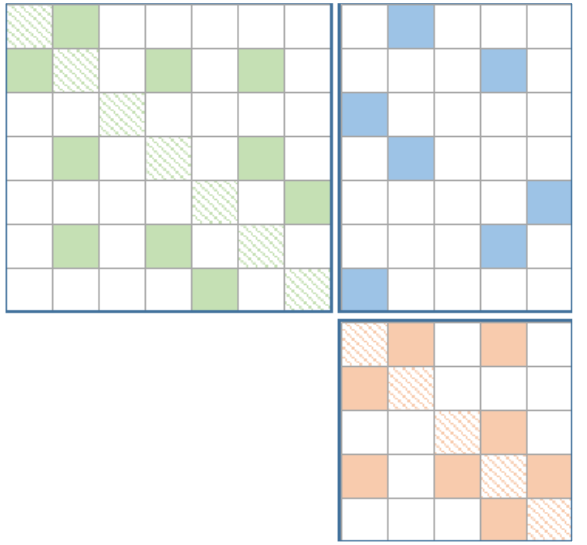






Quality depends on the sparsity of data.

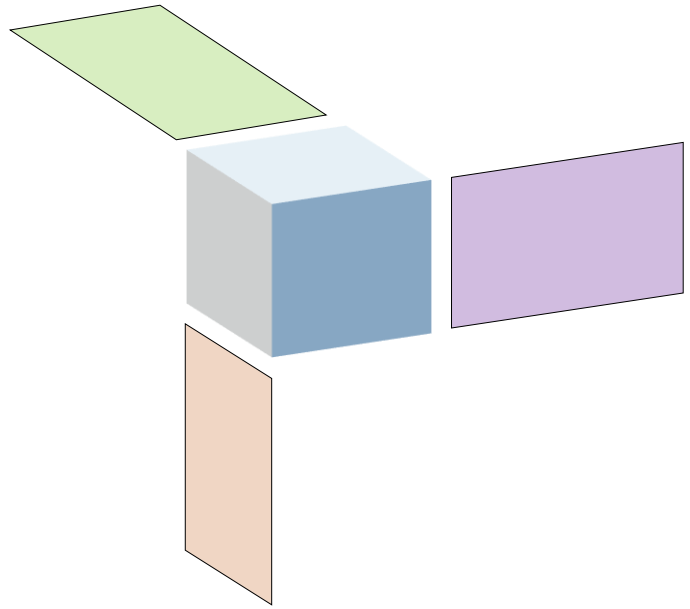
HybridSVD – summary



- ✓Addresses the extreme data sparsity problems by incorporating side information (content data).
- ✓Generates meaningful (more structured) latent feature space.
- ✓Supports quick online recommendations.
- ✓The added complexity is linear w.r.t. the rank of decomposition.
- ✓Applicable in other machine learning areas, e.g., NLP (word embeddings).

- ❑ in the case of rating data can lead to spurious correlations,
- ❑ not (yet) a higher order model.

3. Higher order hybrid model

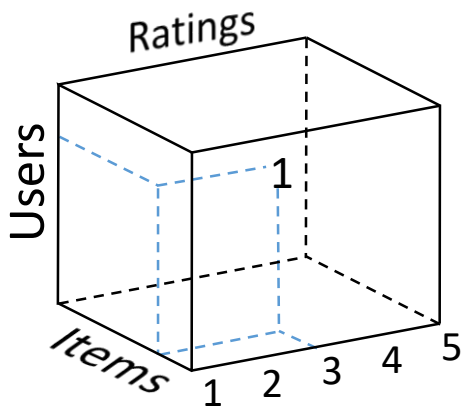


Unified view

- The proposed models address different issues and have their own pitfalls.
- **CoFFee** model is more susceptible to the sparsity issue due to higher order formulation.
- **HybridSVD** may introduce undesired spurious correlations.

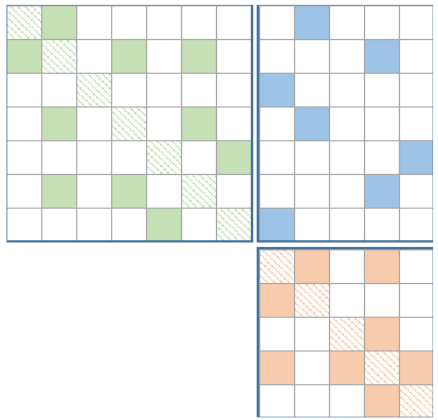
Main idea: combine the previous two methods into a unified approach.

CoFFee model

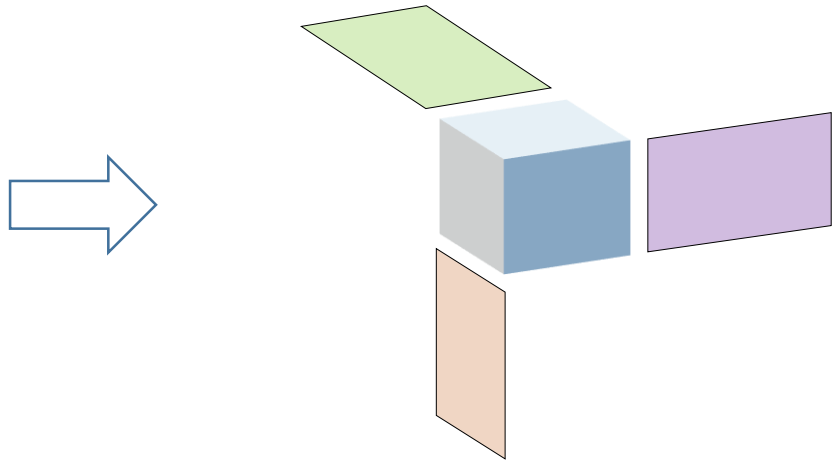


+

HybridSVD



HybridCoFFee



HybridCoFFee

Higher order generalization of HybridSVD.

An auxiliary tensor $\hat{\mathcal{A}}$ can be represented in the form:

$$\hat{\mathcal{A}} \equiv \mathcal{A} \times_1 L_K^T \times_2 L_S^T \times_3 L_R^T, \qquad L_K L_K^T = K, \;\; L_S L_S^T = S, \;\; L_R L_R^T = R.$$

Connection between the auxiliary and the original latent representation:

$$\hat{U} = L_K^T U, \qquad \hat{V} = L_S^T V, \qquad \hat{W} = L_R^T W.$$

Higher order generalization of hybrid folding-in.

Matrix of predicted user preferences for item-context:

$$P = V V_S^T A W_R W^T, \qquad V_S = L_S \hat{V}, \;\; W_R = L_R \hat{W}.$$

Efficient computation

Practical modification of the higher order orthogonal iteration algorithm

Input : Tensor $\mathcal{A}$ in sparse format.
Tensor decomposition ranks r_1, r_2, r_3.
Cholesky factors L_K, L_S, L_R.

Output: auxiliary low rank representation $\mathcal{G}, \hat{U}, \hat{V}, \hat{W}$.

Initialize $\hat{V}, \hat{W}$ *by random matrices with orthonormal columns.*

Compute $V_S = L_S\hat{V},\ W_R = L_R\hat{W}$.

Repeat:

$\hat{U} \leftarrow r_1$ leading left singular vectors of $L_K^T A^{(1)}(W_R \otimes V_S)$,

$U_K \leftarrow L_K\hat{U}$,

$\hat{V} \leftarrow r_2$ leading left singular vectors of $L_S^T A^{(2)}(W_R \otimes U_K)$,

$V_S \leftarrow L_S\hat{V}$,

$\hat{W}, \Sigma, Z \leftarrow r_3$ leading left singular vectors of $L_R^T A^{(3)}(V_S \otimes U_K)$,

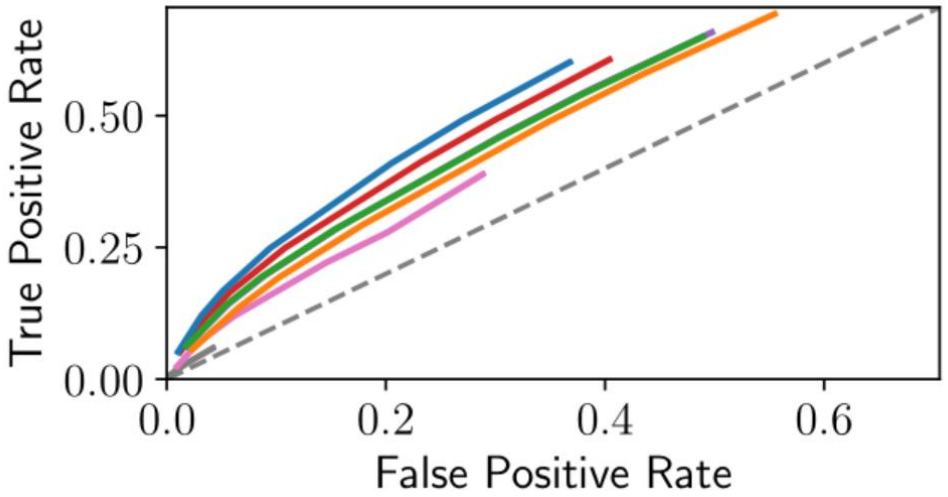
$W_S \leftarrow L_R\hat{W}$,

$\mathcal{G} \leftarrow$ reshape matrix ΣZ^T into shape (r_3, r_1, r_2) and transpose.

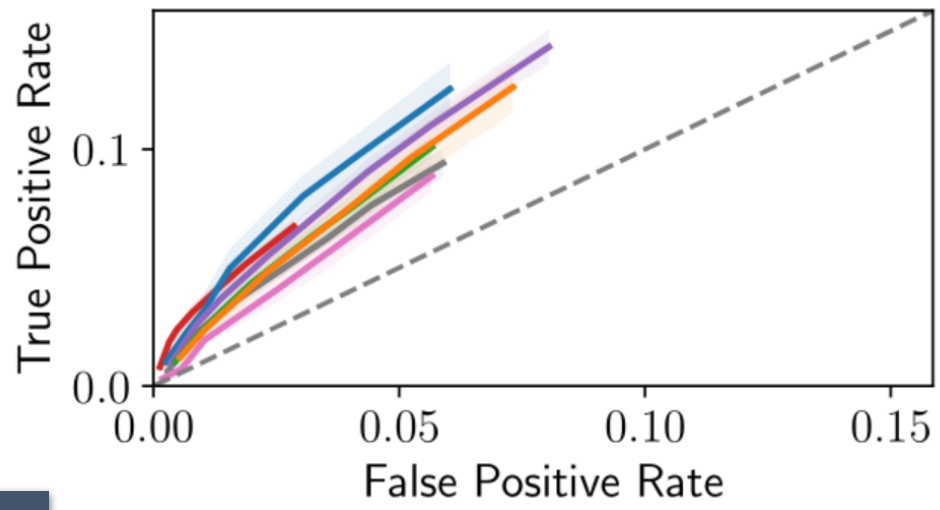
Until: *norm of* $\mathcal{G}$ *ceases to grow or algorithm exceeds maximum number of iterations.*

Results

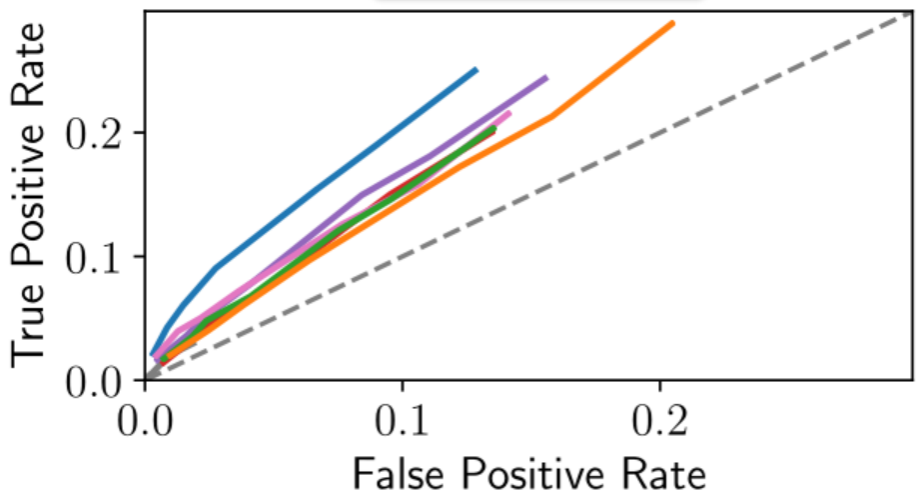
Movielens 10M



BookCrossing



3% fraction of Movielens 10M



Conclusions

- ✓Combines CoFFee model with the HybridSVD approach.
- ✓Efficient computational scheme based on a hybrid modification of a standard HOOI algorithm is proposed.
- ✓Inherits the benefits of its predecessors and at the same time compensates their shortcomings.
- ✓Potentially applicable to a wider class of problems: context-aware, multi-criteria, etc.
- ✓Naturally addresses context vs. content dichotomy.

Directions for future of research:

- ❑ Not feasible for the number of dimensions greater than 4.
- ❑ More appropriate tensor formats (TT/HT) can be used.
- ❑ Tensor-variate normal in application to Gaussian processes?

Thank you!