Clustered Monotone Transforms for Rating Factorization

Gaurush Hiranandani[2*], Raghav Somani[1*], Oluwasanmi Koyejo[2] and Sreangsu Acharyya[1]

[1]Microsoft Research, India and [2]University of Illinois at Urbana-Champaign

Abstract

- **Existing problem**: Existing recommendation engines force users with heterogeneous rating profiles to map their intrinsic rating scales to a common rating scale (e.g. 1-5) of the engine which shatters the low dimensional structure of the rating matrix resulting in a poor fit.
- **Our approach**: We address the (non-linear) scale mismatch between users and the engine by performing regression up to monotonic transformations.
- **Our algorithms**:
  - perform regression up to unknown monotonic transforms over unknown population segments combining the underlying matrix factorization model to exploit the shared low dimensional structure, and
  - have a unique solution in terms of transformed rating scale and regression matrix under verifiable conditions.

Problem Setup

- $\mathcal{U}$ and $\mathcal{V}$ are the set of users and items. $|\mathcal{U}| = N, |\mathcal{V}| = M$. $\mathcal{V}_i$ is the set of items rated by user i.
- $R_{ij} \in [L]$ and $\hat{R}_{ij} \in \mathbb{R}$ are true and predicted ratings of user i to item j. $L \in \mathbb{N}$. $\mathbf{r}^*$: base rating vector $[L, L-1, \ldots, 1]$.
- $W \in \{0,1\}^{N\times M}$, where $w_{ij} = 1$ iff user i rated item j.
- $E_i \in \mathbb{R}^{|\mathcal{V}_i|\times L}$ is the one hot representation of the ratings of user i for each item. $\therefore R_{i,:}^T = E_i\mathbf{r}^*$.

Bregman Divergence

$$D_\phi(\mathbf{x}\,||\,\mathbf{y}) := \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \nabla\phi(\mathbf{y})\rangle$$

ϕ : dom $(\phi) \to \mathbb{R}$ is strictly convex, closed and differentiable on int (dom (ϕ)).

- $D_\phi(\mathbf{x}\,||\,\mathbf{y}) \geq 0$ & $D_\phi(\mathbf{x}\,||\,\mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$. E.g. KL Divergence, Generalized I Divergence, Squared ℓ_2 metric.

Matrix Factorization

Prediction of the form $\hat{R}_{ij} = \langle \mathbf{u}_i, \mathbf{v}_j\rangle$ where $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^d$ are user and item factors respectively.

$$\min_{U,V} \sum_{i\in\mathcal{U}, j\in\mathcal{V}} \frac{1}{2} w_{ij}\left(R_{ij} - \hat{R}_{ij}\right)^2$$

- Enforces a low rank structure on $\hat{R}$; rank $\left(\hat{R}\right) \leq d$.

CMTRF

- CMTRF transforms the base rating scale $\mathbf{r}^*$ of size L.
- Explores partitions of users into groups that share same monotonic transformation of the rating scale.
- Learns the transformed rating non-parametrically from a subset of the set $\mathcal{R}_{L,\epsilon} = \left\{\mathbf{r} \in \mathbb{R}^L \mid r_k \geq r_{k+1} + \epsilon \;\forall\, k \in [L-1], \epsilon > 0\right\}$.

Formulating the optimization problems

If $D(\cdot,\cdot) = D_\phi(\cdot\,||\,\cdot)$ is the loss function for some ϕ, and f is a regression function, we propose 3 variants of the optimization formulation

- **One transformation for all users**:

$$\min_{U,V,\{\mathbf{r}\in\mathcal{R}_{L,\epsilon}\}} \sum_{i\in\mathcal{U}} D\left(E_i\mathbf{r}, f(V_i\mathbf{u}_i)\right) \quad (1)$$

- **Separate transformation for all users**:

$$\min_{U,V,\{\mathbf{r}_i\in\mathcal{R}_{L,\epsilon}\}} \sum_{i\in\mathcal{U}} D\left(E_i\mathbf{r}_i, f(V_i\mathbf{u}_i)\right) \quad (2)$$

- **Separate transformation for each cluster**: Consider K clusters in $\mathcal{U}$

$$\min_{\substack{U,V,\{\mathbf{z}_i\in\{0,1\}^K\}_{i=1}^N,\\ \{\mathbf{r}_k\in\mathcal{R}_{L,\epsilon}\}_{k=1}^K}} \sum_{k\in[K]}\sum_{i\in\mathcal{U}} z_{ik} D\left(E_i\mathbf{r}_k, f(V_i\mathbf{u}_i)\right) \quad (3)$$

where $\mathbf{z}_i$ denotes the one hot encoding for cluster assignment for user i.

<u>**Note**</u>: The monotonic transformations and the clusterings are a-priori unknown and are obtained from the data.

Cost function

- We choose $D(\cdot,\cdot)$ to be a Bregman Divergence $D_\phi(\cdot\,||\,\cdot)$ and the regression function $f = (\nabla\phi)^{-1}$ so that the optimization formulations are tractable.
- The objectives defined in (1)-(3) are convex in $\hat{R}_i = V_i\mathbf{u}_i$, and also in $\mathbf{r}$, $\mathbf{r}_i \;\forall\, i \in [N]$ and $\mathbf{r}_k \;\forall\, k \in [K]$.
- The objective functions are separately convex in $\mathbf{r}$ (or $\mathbf{r}_i$) and $V_i\mathbf{u}_i$ and not jointly convex allowing us to use **coordinate-wise minimization** (alternate minimization) over product of convex sets.
- Under mild conditions it can be shown that (1) and (2) using squared loss recovers the unique solution, whereas we can only guarantee local optimality for (3).

Algorithms

We define $C_\phi(\mathbf{x}\,||\,\mathbf{y}) = D_\phi\left(\mathbf{x}\,||\,(\nabla\phi)^{-1}(\mathbf{y})\right)$ for simplicity, and use alternating minimization to solve the optimization problems (1)-(3).

1-CMTRF

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}\in\mathcal{R}_{L,\epsilon}} \sum_{i\in\mathcal{U}} C_\phi\left(E_i\mathbf{x}\,||\,V_i^t\mathbf{u}_i^t\right)$$
$$U^{t+1}, V^{t+1} = \arg\min_{U,V} \sum_{i\in\mathcal{U}} C_\phi\left(E_i\mathbf{x}^{t+1}\,||\,V_i\mathbf{u}_i\right) + \frac{\lambda_u}{2}\|U\|_F^2 + \frac{\lambda_v}{2}\|V\|_F^2$$

N-CMTRF

$$\mathbf{x}_i^{t+1} = \arg\min_{\mathbf{x}_i\in\mathcal{R}_{L,\epsilon}} \sum_{i\in\mathcal{U}} C_\phi\left(E_i\mathbf{x}_i\,||\,V_i^t\mathbf{u}_i^t\right) \;\forall\, i\in\mathcal{U}$$
$$U^{t+1}, V^{t+1} = \arg\min_{U,V} \sum_{i\in\mathcal{U}} C_\phi\left(E_i\mathbf{x}_i^{t+1}\,||\,V_i\mathbf{u}_i\right) + \frac{\lambda_u}{2}\|U\|_F^2 + \frac{\lambda_v}{2}\|V\|_F^2$$

K-CMTRF

$$\mathbf{z}_i^{t+1} = \arg\min_{k\in[K]} C_\phi\left(E_i\mathbf{x}_k^t\,||\,V_i^t\mathbf{u}_i^t\right) \;\forall\, i\in\mathcal{U}$$
$$\mathbf{x}_k^{t+1} = \arg\min_{\mathbf{x}_k\in\mathcal{R}_{L,\epsilon}} \sum_{i:\mathbf{z}_{ik}^{t+1}=1} C_\phi\left(E_i\mathbf{x}_k\,||\,V_i^t\mathbf{u}_i^t\right) \;\forall\, k \text{ in parallel}$$
$$U^{t+1}, V^{t+1} = \arg\min_{U,V} \sum_{k\in[K]}\sum_{i:\mathbf{z}_{ik}^{t+1}=1} C_\phi\left(E_i\mathbf{x}_k^{t+1}\,||\,V_i\mathbf{u}_i\right) + \frac{\lambda_u}{2}\|U\|_F^2 + \frac{\lambda_v}{2}\|V\|_F^2$$

Baselines for Experiments

- **Baselines models**: (a) regularized Matrix Factorization, (b) LMaFit, (c) Monotonic single index model for Matrix Completion (MMC), and (d) Neural Network Matrix Factorization (NNMF).
- **Baseline datasets**: We consider 7 real world datasets, and 2 synthetic datasets (a) SD-1, and (b) SD-2.
- **Splits**: Datasets having timestamps were split chronologically for training (80%) and validation which is more realistic than uniform random split.

Experiment results

Table 1:Datasets description.

Datasets	Users	Items	Ratings	Density	Split
ML100k	751	1616	82,863	6.83%	Chronological
ML1M	5301	3682	901,851	4.62%	
ML10M	62007	10586	6,950,602	1.06%	
GB	42813	9403	4,729,637	1.17%	
Epinions	77264	150497	808,690	0.007%	
Douban	2999	3000	136891	1.52%	Uniform
Flixster	2307	2945	26173	2.01%	
ML100k_u	943	1650	100000	6.43%	

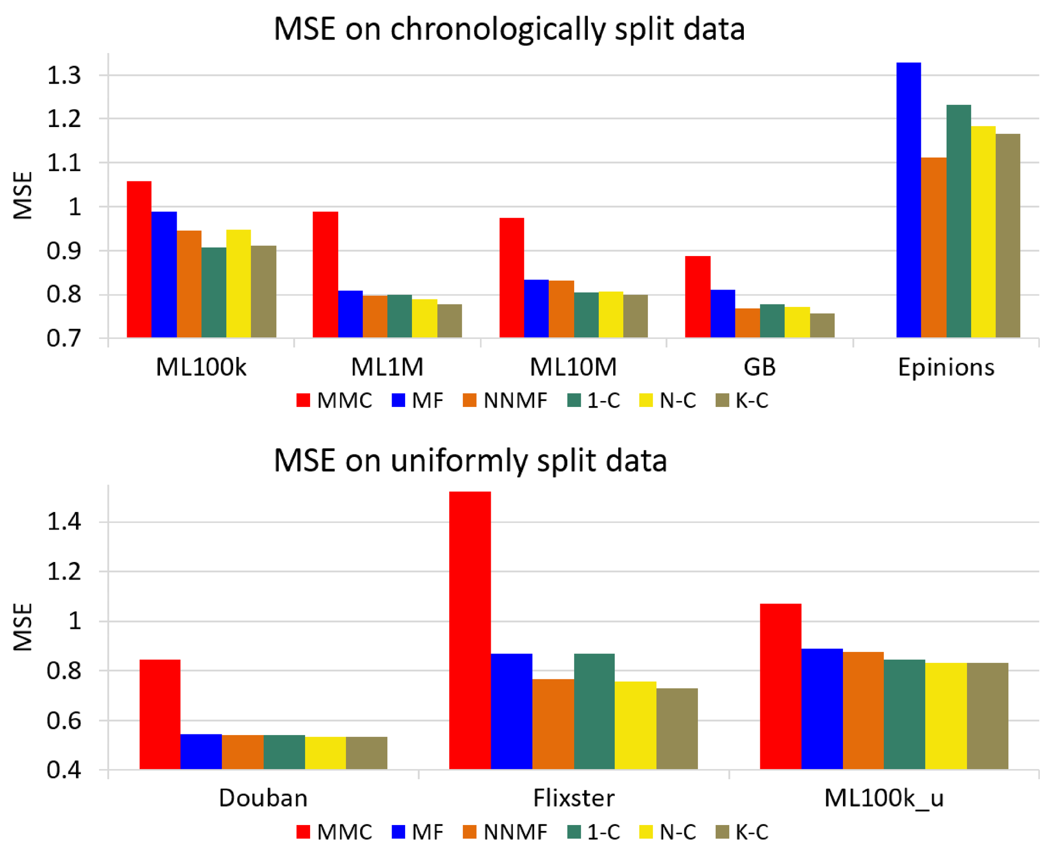

Figure 1:MSE on chronologically and uniformly split test datasets.

Table 2:MSE on synthetic datasets. 1-C, N-C, and K-C denote `1-CMTRF`, `N-CMTRF`, and `K-CMTRF`, respectively.

Data	MF	LMaFit	MMC	NNMF	1-C	N-C	K-C
SD-1	0.140	0.306	0.139	0.137	0.122	**0.122**	0.123
SD-2	0.804	0.674	1.995	1.893	0.326	0.347	**0.326**

*Equal contribution