

Data Context Adaptation for Accurate Recommendation with Additional Information

Hyunsik Jeon
Seoul National University
Seoul, Republic of Korea
jeon185@snu.ac.kr

Bonhun Koo
Seoul National University
Seoul, Republic of Korea
darkgs@snu.ac.kr

U Kang
Seoul National University
Seoul, Republic of Korea
ukang@snu.ac.kr

Abstract—Given a sparse rating matrix and an auxiliary matrix of users or items, how can we accurately predict missing ratings considering different data contexts of entities? Many previous studies proved that utilizing the additional information with rating data is helpful to improve the performance. However, existing methods are limited in that 1) they ignore the fact that data contexts of rating and auxiliary matrices are different, 2) they have restricted capability of expressing independence information of users or items, and 3) they assume the relation between a user and an item is linear.

We propose DACONA, a neural network based method for recommendation with a rating matrix and an auxiliary matrix. DACONA is designed with the following three main ideas. First, we propose a data context adaptation layer to extract pertinent features for different data contexts. Second, DACONA represents each entity with latent interaction vector and latent independence vector. Unlike previous methods, both of the two vectors are not limited in size. Lastly, while previous matrix factorization based methods predict missing values through the inner-product of latent vectors, DACONA learns a non-linear function of them via a neural network. We show that DACONA is a generalized algorithm including the standard matrix factorization and the collective matrix factorization as special cases. Through comprehensive experiments on real-world datasets, we show that DACONA provides the state-of-the-art accuracy.

I. INTRODUCTION

Given a sparse rating matrix and an auxiliary matrix of users or items, how can we accurately predict missing ratings considering different data contexts (e.g., an item belongs to both rating and item-genre contexts) of entities? Predicting unseen rating values is a crucial problem in recommendation because users want to be provided an item list that they will give high ratings.

Matrix factorization (MF) [1]–[3] is a basic yet extensively used method in recommendation due to its simplicity and powerful performance. Given only a sparse rating matrix X , MF derives two low-rank latent matrices U and V that represent user and item features, respectively. It optimizes U and V to reduce the loss $\|X - U^T V\|_F^2$ for observed ratings, where $\|\cdot\|_F^2$ denotes the Frobenius norm. Then MF predicts unobserved rating that user i will give to item j as $U_i^T V_j$, where U_i is i th column of U and V_j is j th column of V . The latent vectors U_i and V_j are trained to represent interaction information. In the real world, however, there exist independence information of users or items which does not directly interact with other information. Biased-MF has been

TABLE I: Comparison of DACONA and other methods. Bold fonts indicate desired settings. DACONA is the only method supporting all the desired properties and providing the richest modeling capability.

Method	Use additional data	Consider data context difference	Model independence information	Linearity
MF [1], [2]	X	X	X	Linear
Biased-MF	X	X	Restricted	Linear
NeuMF [5]	X	X	O	Non-linear
CMF [6]	O	X	X	Linear
Biased-CMF	O	X	Restricted	Linear
FM [7]–[9]	O	X	Restricted	Linear
SREPS [10]	O	O	X	Linear
HybridCDL [11]	O	X	X	Non-linear
DACONA (proposed)	O	O	O	Non-linear

proposed to incorporate such information as well. Biased-MF predicts unobserved rating between user i and item j as $U_i^T V_j + b_i + b_j$, where $b_i \in \mathbb{R}$ and $b_j \in \mathbb{R}$ are bias terms of user i and item j , respectively. However, biased-MF has a restricted model capacity since the inner-product yields only a 1-d scalar value. MF also has a limitation that it models only a linear function. For example, suppose that each dimension of U_i and V_j is trained to represent the degree of each characteristic (e.g., comedy, horror and action in movie genres). MF multiplies the corresponding degrees of characteristics and adds them to predict the rating. In the real world, however, the relationship between a user and an item is not always linear. A user may dislike a movie thoroughly if the degree of fear is below a certain level, regardless of the degree of other characteristics. In order to overcome the limitation of linear model, deep-learning approach is introduced to MF [4], [5]. However, they only utilize a rating matrix although additional information is available in many services.

In recent years, numbers of algorithms have been proposed to use additional information (e.g., social networks [12]–[17], item characteristics [6], [11], [18], and item synopsis [19]–[21]) as well as rating data to improve the performance of recommendation, and they have shown that utilizing both rating data and auxiliary data helps improve the accuracy of rating prediction; thus, effective usage of auxiliary data beyond the rating data has become an important issue in recommendation. We call them as *data context-aware recommendation*, which is different from *context-aware recommendation systems* (CARS) that consider users’ specific situation (e.g., time, place and weather, etc). Collective Matrix Factorization (CMF) [6] is the

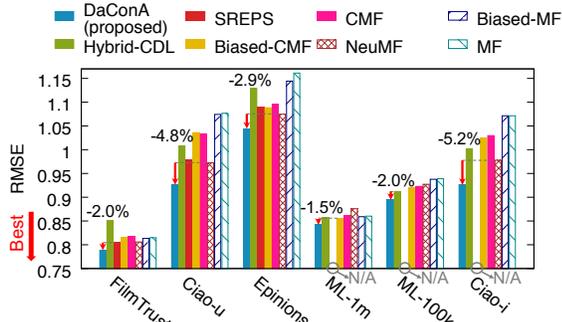


Fig. 1: DAConA shows the best accuracy (lowest RMSE) for six real-world datasets; each percentage indicates the error reduction rate compared to the second best one in each dataset. We omit FM in the plot since it shows far worse performance than the other methods. SREPS is available only for user-coupled datasets.

most popular method for data-context-aware recommendations. CMF comprises two tied MF models which share a single latent matrix and are trained to minimize the losses for a rating matrix and an additional matrix (details in Section II). CMF is also extended to biased-CMF if it is tied to two biased-MF models. However, CMF has a limitation that it directly applies a common latent matrix to two MF models without considering different types of data contexts. For example, assume we have a truster-trustee matrix for auxiliary information, in addition to rating matrix. CMF then forces user latent vectors to represent users’ preferences for both rating and social relationship. However, sharing the same latent vector in different data contexts may not be proper in the real world; users may mainly regard movie’s genre, popularity, and actors to give ratings (rating data context), whereas they consider age, gender, and social status to establish social relationships (social relationship data context). Furthermore, CMF inherits MF’s limitation of predicting the preferences as a linear function and not sufficiently considering independence information.

In this paper, we propose DAConA (Data Context Adaptation for Accurate Rating Recommendation with Additional Information), an accurate recommendation framework based on deep neural networks using both rating matrix and auxiliary data. DAConA is based on the following three main ideas.

- 1) We propose data context adaptation layer in DAConA to adapt latent interaction vector to different data contexts: rating data context and auxiliary data contexts. The data context adaptation layer extracts pertinent features appropriate for different data contexts.
- 2) DAConA represents each entity (e.g., a user and an item) with two types of vectors: *latent interaction vector* and *latent independence vector*. Latent interaction vector is optimized to represent information that is interactive with other entities, whereas latent independence vector is optimized to represent information that is not interactive with other entities. Unlike biased-MF and biased-CMF which use a scalar bias for each entity, latent indepen-

TABLE II: Table of frequently-used symbols.

Symbol	Definition
U, I, C	set of users, items, and context entities
$X \in \mathbb{R}^{ U \times I }$	rating matrix
$Y \in \mathbb{R}^{ I \times C }$	additional matrix for item
Ω_X, Ω_Y	set of indices of observable entries in X and Y
U, V, C	latent interaction matrices of user, item, and additional entity (e.g., trustee and genre)
U^X, V^X	latent independence matrices of user and item in rating data context
V^Y, C^Y	latent independence matrices of item and additional entity in auxiliary data context
D^X, D^Y	data context adaptation matrices for rating data context and auxiliary data context
f_X, f_Y	fully-connected neural networks for X and Y
$\theta_{f_X}, \theta_{f_Y}$	parameters of f_X and f_Y
d_c	dimension of latent interaction vector
d_s	dimension of latent independence vector

dence vector in DAConA is not limited in size.

- 3) We design a neural network framework to predict non-linear relationship between entities. We combine two neural networks using an integrated loss function; they are simultaneously trained to minimize the loss.

Table I compares DAConA with other algorithms in various perspectives. DAConA is the only method that uses additional data, exploits data context difference, models rich independence information, and models non-linear relationship.

The contributions of DAConA are as follows:

- **Algorithm.** We propose DAConA, a method to predict missing values in collective data using neural network. DAConA learns a non-linear function and latent interaction/independence factors, and adapts the latent interaction factors into different contexts via data context adaptation layer.
- **Generalization.** We show that DAConA is a generalized algorithm of well-known MF and Collaborative Filtering algorithms, and further provides rich modeling capability.
- **Performance.** DAConA provides the best accuracy, up to 5.2% lower RMSE than the second best method in real-world datasets (see Figure 1). We also show that our key ideas improve the performance.

Table II lists the symbols used in this paper. Our source code and datasets are available at <https://datalab.snu.ac.kr/dacona/>.

II. PRELIMINARIES: COLLECTIVE MATRIX FACTORIZATION

Traditional Matrix Factorization (MF) suffers from a data sparsity problem [22]; Collective Matrix Factorization (CMF) [6] has been proposed to solve the problem by using additional data. While MF decomposes a single matrix into two latent matrices, CMF decomposes both rating matrix and additional matrix into three latent matrices, and minimizes an integrated loss function which is based on *inner-product*. For a factorization problem of users and items, either user context matrix or item context matrix is available as an additional matrix. Assuming the auxiliary matrix is item-coupled, given a rating matrix $X \in \mathbb{R}^{|U| \times |I|}$ and an auxiliary

matrix $Y \in \mathbb{R}^{|\mathbb{I}| \times |\mathbb{C}|}$, rank- d decomposition of these two data matrices yields three latent matrices, $U \in \mathbb{R}^{d \times |\mathbb{U}|}$, $V \in \mathbb{R}^{d \times |\mathbb{I}|}$ and $C \in \mathbb{R}^{d \times |\mathbb{C}|}$. \mathbb{U} , \mathbb{I} , and \mathbb{C} indicate set of users, items, and auxiliary entities, respectively. V is shared for the predicted matrices \hat{X} and \hat{Y} as follows:

$$\hat{X}_{ij} = U_i^T V_j, \quad \hat{Y}_{jk} = V_j^T C_k,$$

where U_i , V_j , and C_k are i th column of U , j th column of V , and k th column of C , respectively. U , V , and C are trained to minimize the following loss function:

$$L = \frac{1}{2} \sum_{(i,j) \in \Omega_X} (\hat{X}_{ij} - X_{ij})^2 + \frac{1}{2} \sum_{(j,k) \in \Omega_Y} (\hat{Y}_{jk} - Y_{jk})^2 + \frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2 + \|C\|_F^2),$$

where Ω_X and Ω_Y are sets of indices of observable entries in X and Y , respectively. λ is a regularization parameter to prevent overfitting. CMF is extended to biased-CMF if bias terms are added. Biased-CMF predicts ratings and entries of auxiliary matrix as follows:

$$\hat{X}_{ij} = U_i^T V_j + b_i + b_j, \quad \hat{Y}_{jk} = V_j^T C_k + \bar{b}_j + \bar{b}_k,$$

where $b_i, b_j, \bar{b}_j, \bar{b}_k \in \mathbb{R}$ are bias terms. The bias terms represent independence information of entities.

CMF based method [12] shows better performance than MF [1], [2] for real-world datasets. However, CMF has limitations that it predicts the relationship between entities by linear functions and does not consider the fact that data context of the rating matrix and that of the additional matrix are different. Moreover, even biased-CMF is not able to sufficiently consider independence information because the dimensions of all bias terms are restricted: each entity has a scalar bias term.

III. PROPOSED METHOD

We describe our proposed method DACONA for rating prediction. After presenting the overview of our method in Section III-A, we formulate the objective function of DACONA in Section III-B. Then we describe three key ideas of DACONA in Sections III-C, III-D, and III-E, respectively. Afterwards, we explain the training algorithm of DACONA in Section III-F. Finally, we compare DACONA with existing methods in Section III-G showing that DACONA is a generalized version of them.

A. Overview

DACONA utilizes a rating matrix and an auxiliary matrix to predict unseen ratings. We design DACONA based on the following observations: 1) there exist independence factors such as biases as well as interaction factors that affect the relationship between entities in real-world, 2) the information needed for different data contexts is different, and 3) users and items do not interact in a simple linear way. Based on these observations, the goals of DACONA are 1) to learn embedding vectors that represent interaction and independence features of entities, 2) to extract appropriate features for each data context from a shared latent factor, and 3) to learn complex

non-linear scoring function between entities. Designing such recommendation model entails the following challenges:

- 1) **Modeling entities.** How can we model the embedding vector for each entity to represent interaction and independence features?
- 2) **Considering data context difference.** How can we learn the common factor between rating matrix and an auxiliary matrix (e.g., item-genre matrix or truster-trustee matrix) even though they are in different data contexts?
- 3) **Non-linear relationship.** How can we learn a scoring function that grasps non-linear relationship between entities?

We propose the following main ideas to address the challenges:

- 1) **Latent interaction and independence factors.** We map each entity into two types of latent factors: *latent interaction factor* which captures interaction features between entities, and *latent independence factor* which captures independence features.
- 2) **Data context adaptation.** We adapt the latent interaction vectors to each data context through data context adaptation layer before scoring the relationship between two entities.
- 3) **Non-linear modeling.** Fully-connected neural networks are trained to score between entities using latent independence factors, and latent interaction factors that are adapted to each data context. Neural networks are capable of seizing non-linear relationship between features because of non-linear activation functions.

Figure 2 shows the architecture of DACONA. Latent interaction vectors (U_i , V_j , and C_k) are projected into each data context by multiplication with data context adaptation matrices (D^X and D^Y). We perform element-wise product on the projected latent interaction vectors to capture interactions between two entities. We then concatenate them with latent independence vectors of entities and feed them into fully-connected neural networks. The networks predict the entries of rating matrix (\hat{X}_{ij}) and the entries of auxiliary matrix (\hat{Y}_{jk}). In the following sections, we formulate DACONA's objective function, describe DACONA's predictive model in details, and compare DACONA with MF and CMF showing that they are limited settings of DACONA. For simplicity, we assume that the auxiliary matrix Y is an items' auxiliary matrix representing genre data context (containing item and genre) in the remainder of this section; it is trivial to use users' auxiliary matrix such as truster-trustee matrix instead. Extending DACONA to utilize multiple additional information, and its experimental results are discussed in Appendix A.

B. Objective Formulation

Let $X \in \mathbb{R}^{|\mathbb{U}| \times |\mathbb{I}|}$ be a rating matrix and $Y \in \mathbb{R}^{|\mathbb{I}| \times |\mathbb{C}|}$ be an items' auxiliary matrix such as item-genre matrix, where \mathbb{U} , \mathbb{I} , and \mathbb{C} indicate sets of users, items, and auxiliary entities,

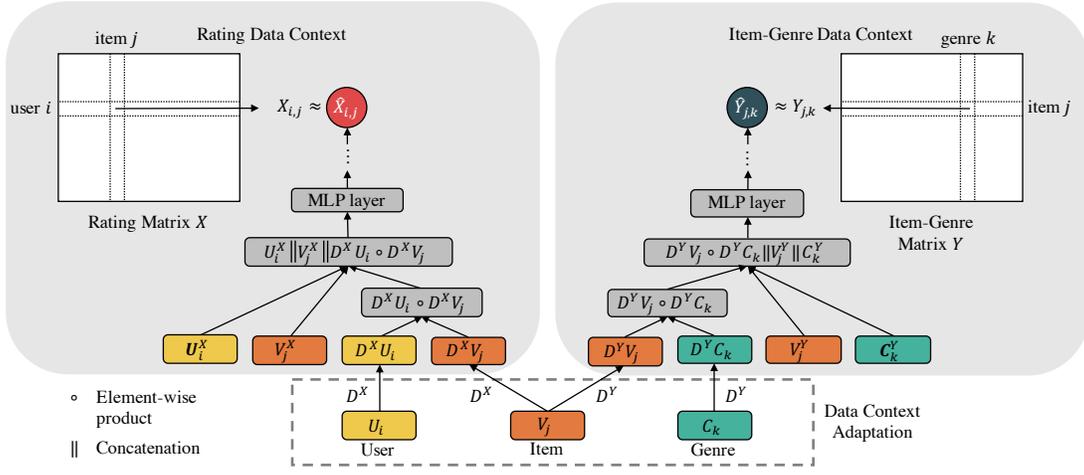


Fig. 2: Architecture of DACONA for item coupled dataset.

respectively. We then formulate the objective function (to be minimized) of DACONA as follows:

$$L = (1 - \alpha)loss_X + \alpha loss_Y, \quad (1)$$

where

$$loss_X = \frac{1}{2} \sum_{(i,j) \in \Omega_X} (\hat{X}_{ij} - X_{ij})^2 + \frac{\lambda}{2} Reg_X, \quad (2)$$

$$loss_Y = \frac{1}{2} \sum_{(j,k) \in \Omega_Y} (\hat{Y}_{jk} - Y_{jk})^2 + \frac{\lambda}{2} Reg_Y. \quad (3)$$

Ω_X and Ω_Y are sets of indices of observable entries in X and Y , respectively. \hat{X}_{ij} and \hat{Y}_{jk} are predicted entries from DACONA. Reg_X and Reg_Y are regularization terms controlled by λ to prevent overfitting (details in Section III-E). $0 \leq \alpha \leq 1$ is a hyperparameter that controls the balance between $loss_X$ and $loss_Y$.

C. Latent Interaction/Independence Factors

Latent interaction vector represents the interaction information of entities. Let $U \in \mathbb{R}^{d_c \times |\mathbb{U}|}$, $V \in \mathbb{R}^{d_c \times |\mathbb{I}|}$, and $C \in \mathbb{R}^{d_c \times |\mathbb{C}|}$ be latent interaction matrices of users, items, and additional entities, respectively, where d_c is the dimension of latent interaction vector. Then i th column $U_i \in \mathbb{R}^{d_c}$ of U , j th column $V_j \in \mathbb{R}^{d_c}$ of V , and k th column $C_k \in \mathbb{R}^{d_c}$ of C are user i 's, item j 's, and additional entity k 's latent interaction vector, respectively, which are optimized through training process. Assume that \mathbb{I} is the set of movies. If user i likes romance movies and movie j is about romance, the relationship between the two entities has to be strong. Conversely, if user i likes comic movies but movie j is not comic at all, the relationship between the two entities has to be weak. Matrix Factorization (MF) captures interaction features of entities by performing inner-product of two latent interaction vectors; DACONA generalizes MF by performing element-wise product of two latent interaction vectors, and applies non-linear neural network layers on top of them. Note that the latent interaction vectors are adapted to each data context as described in Section III-D; however, the concept that the latent interaction vectors capture the interaction information does not change.

Latent independence vector represents the independence information for each entity i , j , and k . If item i is of very high quality, then the item will receive high ratings from most users, regardless of their tastes. Therefore, we need latent independence vectors to represent independence information which latent interaction vectors cannot represent. Also, an item may receive high ratings from many users in rating data context, while the item does not belong to many genres in the genre data context. To address such issue, we model each entity to have separate latent independence vector for each data context. Let $U^X \in \mathbb{R}^{d_s \times |\mathbb{U}|}$ and $V^X \in \mathbb{R}^{d_s \times |\mathbb{I}|}$ be latent independence matrices for users and items, respectively, in rating data context. Let $V^Y \in \mathbb{R}^{d_s \times |\mathbb{I}|}$ and $C^Y \in \mathbb{R}^{d_s \times |\mathbb{C}|}$ be latent independence matrices for items and genres, respectively, in item-genre data context. Then $U_i^X \in \mathbb{R}^{d_s}$ and $V_j^X \in \mathbb{R}^{d_s}$ are user i 's and item j 's latent independence vectors respectively in rating data context. Similarly, $V_j^Y \in \mathbb{R}^{d_s}$ and $C_k^Y \in \mathbb{R}^{d_s}$ are item j 's and genre k 's latent independence vectors respectively in item-genre data context. All of the latent independence vectors are optimized through training process. Note that the dimension of latent independence vector of DACONA is not limited in size, unlike biased-MF and biased-CMF which is limited to a scalar bias (1-d value).

Formally, the predictive model of DACONA with latent interaction vectors and latent independence vectors is defined as follows:

$$\begin{aligned} \hat{X}_{ij} &= f_X(T_X(U_i) \circ T_X(V_j), U_i^X, V_j^X), \\ \hat{Y}_{jk} &= f_Y(T_Y(V_j) \circ T_Y(C_k), V_j^Y, C_k^Y), \end{aligned} \quad (4)$$

where T_X and T_Y denote data context adaptation functions for X and Y , respectively (details in Section III-D), f_X and f_Y denote neural networks to predict entries in X and Y , respectively (details in Section III-E), and symbol \circ denotes element-wise product operation. We combine two transferred latent interaction vectors (e.g., $T_X(U_i)$ and $T_X(V_j)$) via element-wise product to learn correlated information.

D. Context Adaptation

To capture features appropriate for each data context from latent interaction vectors, we need to learn data context

adaptation functions (T_X and T_Y) which are in Eq. (4). We propose a data context adaptation layer which projects the latent interaction vectors into each data context with learnable projection matrices; the context adaptation layer could be replaced with any projection function such as neural networks. Let $D^X \in \mathbb{R}^{d'_c \times d_c}$ be a projection matrix for rating data context and $D^Y \in \mathbb{R}^{d'_c \times d_c}$ be a projection matrix for item-genre data context. Then latent interaction vectors for user i and item j are adapted into rating data context as $D^X U_i$ and $D^X V_j$, respectively. Here, we apply a single projection matrix D_X for both user and item latent interaction vectors since we want to learn a general mapping to the rating data context from the two interaction vectors, not a separate mapping from each entity; we experimentally show the effectiveness of sharing the same projection matrix for a data context in Section IV-D. Similarly, we adapt latent interaction vectors of item j and genre k to item-genre data context as $D^Y V_j$ and $D^Y C_k$, respectively. Then Eq. (4) is changed to the following.

$$\begin{aligned}\hat{X}_{ij} &= f_X(D^X U_i \circ D^X V_j, U_i^X, V_j^X), \\ \hat{Y}_{jk} &= f_Y(D^Y V_j \circ D^Y C_k, V_j^Y, C_k^Y).\end{aligned}\quad (5)$$

E. Non-linear Modeling

We introduce fully-connected neural networks f_X and f_Y to deal with non-linear relationships between features. We combine the input features of the networks by concatenating them as follows:

$$\hat{X}_{ij} = f_X\left(\begin{bmatrix} D^X U_i \circ D^X V_j \\ U_i^X \\ V_j^X \end{bmatrix}\right), \hat{Y}_{jk} = f_Y\left(\begin{bmatrix} D^Y V_j \circ D^Y C_k \\ V_j^Y \\ C_k^Y \end{bmatrix}\right), \quad (6)$$

where $f_X : \mathbb{R}^{d'_c + 2d_s} \rightarrow \mathbb{R}$ and $f_Y : \mathbb{R}^{d'_c + 2d_s} \rightarrow \mathbb{R}$ are fully-connected neural networks for rating data context and auxiliary data context, respectively. The square bracket $[\]$ denotes concatenation of vectors. We apply rectifier (ReLU), sigmoid, and hyperbolic tangent (Tanh) as the activation functions in f_X and f_Y from which we observe the followings: 1) although ReLU shows good performance in predicting rating, the variation in the process of convergence is large, 2) sigmoid shows little variation in convergence, but does not show good performance in predicting ratings and the speed of convergence is too slow, and 3) Tanh shows the fastest convergence, little variation, and the best performance. As a result, we use Tanh as an activation function in DACONA for experiments.

We use L_2 -regularization to prevent overfitting. Then Reg_X and Reg_Y respectively in Eq.(2) and Eq.(3) are formulated as follows:

$$\begin{aligned}Reg_X &= \sum_{i \in \mathcal{U}} (\|U_i^X\|^2 + \|D^X U_i\|^2) \\ &+ \sum_{j \in \mathcal{I}} (\|V_j^X\|^2 + \|D^X V_j\|^2) + \sum \|\theta_{f_X}\|^2,\end{aligned}\quad (7)$$

$$\begin{aligned}Reg_Y &= \sum_{j \in \mathcal{I}} (\|V_j^Y\|^2 + \|D^Y V_j\|^2) \\ &+ \sum_{k \in \mathcal{C}} (\|C_k^Y\|^2 + \|D^Y C_k\|^2) + \sum \|\theta_{f_Y}\|^2,\end{aligned}\quad (8)$$

where θ_{f_X} and θ_{f_Y} denote the set of parameters of f_X and f_Y , respectively.

Algorithm 1: Training DACONA

Input : Rating matrix $X \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$, auxiliary matrix $Y \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{C}|}$, set of indices of observable entries Ω_X and Ω_Y , balance parameter α , and regularizer λ

Output: latent interaction matrices U, V , and C , latent independence matrices U^X, V^X, V^Y , and C^Y , and parameters of neural networks θ_{f_X} , and θ_{f_Y}

- 1 Initialize $U, V, C, U^X, V^X, V^Y, C^Y, \theta_{f_X}$, and θ_{f_Y} ;
- 2 **while** *stopping condition is not met* **do**
- 3 **for** $(i, j) \in \Omega_X$ **do**
- 4 $\hat{X}_{ij} \leftarrow$ Eq. (6);
- 5 $Reg_X \leftarrow$ Eq. (7);
- 6 $loss_X \leftarrow$ Eq. (2) with \hat{X}_{ij} , Reg_X , and λ ;
- 7 $loss \leftarrow (1 - \alpha)loss_X$;
- 8 alternately update U, V, U^X, V^X , and θ_{f_X} to minimize $loss$ while fixing the others;
- 9 **end for**
- 10 **for** $(j, k) \in \Omega_Y$ **do**
- 11 $\hat{Y}_{jk} \leftarrow$ Eq. (6);
- 12 $Reg_Y \leftarrow$ Eq. (8);
- 13 $loss_Y \leftarrow$ Eq. (3) with \hat{Y}_{jk} , Reg_Y , and λ ;
- 14 $loss \leftarrow \alpha loss_Y$;
- 15 alternately update V, C, V^Y, C^Y , and θ_{f_Y} to minimize $loss$ while fixing the others;
- 16 **end for**
- 17 **end**

F. Training DACONA

Algorithm 1 shows how DACONA predicts the values in X and Y , and optimizes the parameters. We first initialize all parameters via Xavier normalization [23] (line 1). Then we draw samples from Ω_X (line 3) and Ω_Y (line 10), respectively, and learn parameters for $loss_X$ (lines 4-8) and $loss_Y$ (lines 11-15) alternately. Specifically, we iteratively optimize a parameter block while fixing the remaining parameter blocks (lines 8, 15); the derivative of each parameter is calculated with back-propagation. We repeat the update procedure until the validation error converges (line 2). We adopt Adaptive Moment Estimation (*Adam*) [24] which is a first-order gradient-based optimizer for updating all parameters. The hyperparameters (d_c, d_s, l, λ , and α) are tuned by grid-search through validation tests. The optimal settings of hyperparameters are presented in Section IV-A.

G. Generality of DACONA

We show that DACONA is a generalized algorithm of several existing methods.

- **MF** [1], [2]. DACONA with the following parameters equals to MF: 1) setting $\alpha = 0$, 2) setting $d'_c = d_c$ and D^X to identity matrix, 3) setting $d_s = 0$, which means there exist no latent independence vectors, 4) setting the number of layers of f_X to 1 and all weights of the layer to 1, which means the network merely sums all input values, and 5) removing activation function, which means that the model captures only linear features.
- **Biased-MF**. MF is extended to biased-MF if bias term

TABLE III: Dataset statistics.

Dataset name	Auxiliary data type	Rating matrix				Auxiliary matrix			
		# observed ratings	# users	# items	density	# observed entries	# rows	# columns	density
Epinions ¹	user information	658,621	44,434	139,374	0.01%	487,183	44,434	49,288	0.02%
Ciao-u ²	user information	72,665	18,133	16,121	0.02%	40,133	18,133	4,299	0.05%
FilmTrust ³	user information	35,093	1,461	2,067	1.16%	1,853	1,461	732	0.17%
ML-1m ⁴	item information	1,000,209	6,040	3,883	4.26%	73,777	3,883	19	100%
ML-100k ⁵	item information	100,000	943	1,682	6.30%	31,958	1,682	19	100%
Ciao-i ²	item information	72,655	18,133	16,121	0.02%	274,057	16,121	17	100%

of each entity is added in the loss function. Therefore, if we set $d_s = 1$ (at step 3) in the setting of MF above, DACONA becomes biased-MF.

- **CMF** [6]. CMF is composed of two MF models and it has the following settings of DACONA: 1) setting $d_c'' = d_c'$, and D^X and D^Y to identity matrix, 2) setting $d_s = 0$, 3) setting the number of layers of f_X and f_Y to 1 and all weights of f_X and f_Y as 1, and 4) removing activation function.
- **Biased-CMF**. CMF is extended to biased-CMF by adding bias term of each entity in the loss function. Thus, if we set $d_s = 1$ (at step 2) in the setting of CMF above, DACONA becomes biased-CMF.

IV. EXPERIMENTS

We perform experiments to answer the following questions.

- **Q1. (Overall performance)** How better is DACONA compared to competitors? (Section IV-B)
- **Q2. (Effects of interaction and independence factors)** How do dimensions of *latent interaction and independence vectors* affect the performance? (Section IV-C)
- **Q3. (Effects of data context adaptation)** How does data context adaptation layer affect the performance? (Section IV-D)
- **Q4. (Neural networks)** Do deeper structures yield better performance? Is activation function helpful to improve the performance? (Section IV-E)

A. Experimental Setup

Metrics. We use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics.

$$\text{RMSE} = \sqrt{\frac{\sum_u \sum_i (\hat{X}_{ui} - X_{ui})^2}{|\text{test ratings}|}}, \text{MAE} = \frac{\sum_u \sum_i |\hat{X}_{ui} - X_{ui}|}{|\text{test ratings}|}.$$

Smaller RMSE and MAE indicate better performance since it means the predicted value is closer to the real value.

Datasets. We use real-world datasets Epinions, Ciao, FilmTrust, MovieLens-1m (ML-1m), and MovieLens-100k (ML-100k) summarized in Table III. We call Ciao dataset as Ciao-i when using items' auxiliary data, and as Ciao-u when using users' auxiliary data. Epinions, Ciao-u, and FilmTrust provide users' auxiliary information (user-user trust

relationship), while ML-1m, ML-100k, and Ciao-i provide items' auxiliary information (movie-genre). For each dataset, 80% of rating data are randomly selected for training and the rest are used for test. 98% of training set are used for training models, and the rest are used for validation in training process. All of the auxiliary data are used only for training since our goal is not to predict entries in the auxiliary matrix.

Competitors. We compare DACONA to the following state-of-the-art algorithms. Comparisons of DACONA and the competitors are summarized in Table I.

- MF [1], [2]. MF factorizes a rating matrix into user and item factors which are combined via inner-product.
- Biased-MF. 1-dimensional bias variables are introduced to MF to model independence factors of entities.
- NeuMF [5]. NeuMF uses a neural network to capture non-linear relationship between entities.
- CMF [6]. CMF is composed of two MFs with an integrated objective function; it learns three low rank matrices by factorizing both rating matrix and auxiliary matrix simultaneously.
- Biased-CMF. 1-dimensional bias variables for entities are introduced to CMF.
- FM [7]–[9]. FM factorizes a rating matrix to user, item, and auxiliary entity factors. We use a 2-way FM which captures all single and pairwise interactions.
- SREPS [10]. SREPS is composed of matrix factorization model, recommendation network model, and social network model while sharing user latent matrix between the models through separate projections for each model.
- Hybrid-CDL [11]. Hybrid-CDL consists of two Additional Stacked Denoising Autoencoders (aSDAE), one is for users and the other for items. Each aSDAE encodes rating vector of an entity and auxiliary data vector of the entity to a latent vector. Since we are given one auxiliary matrix with rating matrix, we adopt aSDAE for entities (e.g., item) *with* additional information and adopt SDAE (Stacked Denoising Autoencoders) for entities (e.g., user) *without* additional information.

Hyperparameters. We find the optimal hyperparameters via grid search for all the competitors and DACONA. The optimal hyperparameters for each method are shown in Table V. We set $d_c = d_c' = d_c''$ in all experiments for simplicity. We use *tower structure* for f_X and f_Y which is widely used for fully-connected networks [5], [25]. In the tower structure, dimension of a layer is half of that of the previous layer. We set the structures of f_X and f_Y to (40, 20, 10) for all experiments. For fair comparisons, we set the dimensions of predictive factors

¹http://www.trustlet.org/downloaded_epinions.html

²<https://www.librec.net/datasets/CiaoDVD.zip>

³<https://www.librec.net/datasets/filmtrust.zip>

⁴<https://grouplens.org/datasets/movielens/1m>

⁵<https://grouplens.org/datasets/movielens/100k>

TABLE IV: DAConA provides the best accuracy in rating prediction. Bold text and * indicate the lowest and the second lowest errors, respectively. The last column denotes the error reduction rate of DAConA compared to the second best one.

Datasets	Metrics	MF	Biased-MF	NeuMF	CMF	Biased-CMF	FM	SREPS	Hybrid-CDL	DAConA (proposed)
Epinions (user-coupled)	RMSE	1.1612	1.1443	1.0746*	1.0956	1.0883	1.1557	1.0887	1.1293	1.0433 (-2.9%)
	MAE	0.8965	0.8845	0.8270*	0.8513	0.8423	0.8977	0.8462	0.8863	0.7996 (-3.3%)
Ciao-u (user-coupled)	RMSE	1.0765	1.0742	0.9724*	1.0568	1.0451	1.5734	0.9786	1.0077	0.9255 (-4.8%)
	MAE	0.8349	0.8352	0.7813	0.8195	0.8127	1.2018	0.7692*	0.7851	0.6894 (-10.4%)
FilmTrust (user-coupled)	RMSE	0.8148	0.8133	0.8058	0.8171	0.8153	0.9053	0.8040*	0.8504	0.7882 (-2.0%)
	MAE	0.6325	0.6334	0.6240	0.6320	0.6270	0.6769	0.6168*	0.6709	0.6035 (-2.2%)
ML-1m (item-coupled)	RMSE	0.8601	0.8590	0.8758	0.8561	0.8553*	0.8843	N/A	0.8569	0.8422 (-1.5%)
	MAE	0.6723	0.6715	0.6921	0.6701	0.6691*	0.6861	N/A	0.6715	0.6597 (-1.4%)
ML-100k (item-coupled)	RMSE	0.9389	0.9378	0.9273	0.9227	0.9200	0.9286	N/A	0.9118*	0.8938 (-2.0%)
	MAE	0.7415	0.7396	0.7248	0.7234	0.7210	0.7330	N/A	0.7156*	0.6988 (-2.3%)
Ciao-i (item-coupled)	RMSE	1.0713	1.0672	0.9777*	1.0296	1.0238	1.3387	N/A	1.0019	0.9264 (-5.2%)
	MAE	0.8330	0.8293	0.7778*	0.8133	0.8130	1.0478	N/A	0.7830	0.7088 (-8.9%)

TABLE V: Hyperparameters for each dataset. There are no hyperparameters of SREPS for ML-1m, ML-100k, and Ciao-i since SREPS is not available for item-coupled datasets.

Method	Hyper-parameters	Epinions/Ciao-u/FilmTrust/ML-1m/ML-100k/Ciao-i						
MF, Biased-MF	l	1e-3	1e-3	1e-3	1e-3	1e-4	1e-4	
	λ	1e-5	1e-4	1e-3	1e-5	1e-4	1e-4	
NeuMF	l	1e-3	1e-4	1e-3	1e-3	1e-3	1e-3	
	λ	1e-5	1e-5	1e-4	1e-5	1e-5	1e-4	
CMF, Biased-CMF	l	1e-3	1e-4	1e-3	1e-4	5e-4	1e-4	
	λ	1e-5	1e-4	1e-4	1e-5	1e-5	1e-5	
FM	l	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	
SREPS	l	1e-3	1e-3	1e-3	N/A	N/A	N/A	
	λ	5e-6	1e-5	1e-4	N/A	N/A	N/A	
	α	0.4	0.3	0.2	N/A	N/A	N/A	
	β	0.1	0.2	0.1	N/A	N/A	N/A	
Hybrid-CDL	l	1e-3	1e-3	1e-3	1e-4	1e-3	1e-3	
	λ	1e-2	1e-2	1e-4	1e-3	1e-3	1e-3	
	α_1	0.8	0.2	0.5	0.2	0.2	0.2	
	α_2	0.2	0.8	0.2	0.2	0.5	0.8	
DAConA (proposed)	l	1e-3	1e-4	1e-3	1e-3	1e-3	1e-3	
	λ	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	
	α	0.4	0.8	0.9	0.2	0.9	0.6	
	d_s	13	14	11	7	10	11	

of all methods to 10; predictive factor indicates the factor that directly decides the ratings. For instances, predictive factors of each method are as follows: user/item latent vectors of MF, CMF, FM and SREPS, the last hidden layer of NeuMF and DAConA, and the encoded latent factors of hybrid-CDL. For NeuMF, we assign 5-dimension of the predictive factor to GMF and the other 5-dimension of that to MLP; then we use tower structure for the MLP and set α to 0.5 as in [5]. For hybrid-CDL, we adopt [500, 10, 500]-structured hidden layers to aSDAE/SDAE; we set the first dimension of the hidden layers to 500 since it gives the best results among {1000, 500, 250, 100} for all experiments. For biased-MF and biased-CMF, we set the dimensions of user/item latent vectors to 8 since two 1-dimensional bias variables are also used for prediction.

B. Performance Comparison (Q1)

We measure rating prediction errors of DAConA and the competitors in six real-world datasets. Table IV shows RMSEs and MAEs of DAConA and those of the competitors; SREPS is only available in user-coupled datasets since it utilizes the auxiliary matrix through network embedding. Note that

DAConA consistently outperforms competitors in rating prediction for all six real-world datasets. Among the competitors, models that use additional information (e.g., CMF) usually perform better than models using only rating information (e.g., MF), suggesting the importance of utilizing the additional information. Models considering the latent independence factor (e.g., biased-MF and biased-CMF) show better performances than models without the latent independence factor (e.g., MF and CMF) on all datasets, suggesting the importance of considering the independence information. SREPS outperforms CMF on all user-coupled datasets, suggesting the effectiveness of sharing embedding vectors for different data contexts via an indirect way such as projection. However, SREPS does not share the projection matrices between the entities in a data context; DAConA shares projection matrix between entities in a data context which improves the accuracy significantly (details in Section IV-D). NeuMF outperforms biased-MF, suggesting the effectiveness of non-linear modeling via neural networks. However, NeuMF captures non-linear features only from latent independence factors but not from latent interaction factors; DAConA captures non-linear features from not only latent independence factors but also latent interaction factors since interaction factors also affect the relationship between entities non-linearly. Hybrid-CDL does not perform well for user-coupled datasets which include extremely sparse auxiliary matrices since it directly feeds sparse auxiliary vectors (e.g., $Y_{j,:}$) into the model. A supplementary experiment is in Appendix A where we compare the performance of DAConA and hybrid-CDL given two additional information (item-genre matrix and user-trustee matrix).

C. Effects of Interaction and Independence Factors (Q2)

We empirically verify the effectiveness of balancing the dimensions of latent interaction and independence factors on real-world datasets in Figure 3. We measure RMSE of DAConA while varying the dimension d_s of latent independence vector. The dimension d_c of latent interaction vector is determined to be $(40 - 2d_s)$ to maintain the network structure, and the other hyperparameters are fixed as optimal settings which are reported in Table V. Increasing d_s enlarges the model capacity to embed independence information, while it reduces the capacity to embed interaction information. We

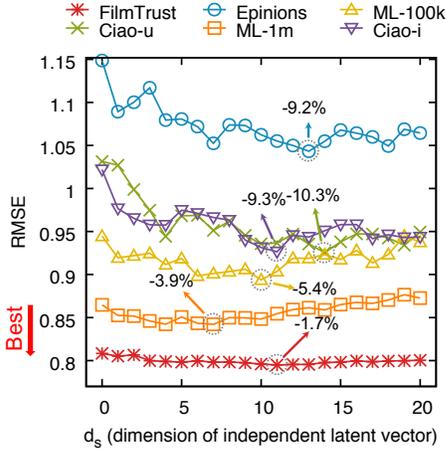


Fig. 3: Considering both interaction and independence information is crucial. The plot shows change of RMSE varying the dimension d_s of latent independence vector. Each percentage indicates the reduction rate of the best RMSE compared to the worst RMSE in each dataset. Note that the best RMSE is achieved when d_s is not too small or too large.

train and test the model for each setting ten times and average them. RMSEs are the lowest when d_s is set to the followings: (11 in FilmTrust), (14 in Ciao-u), (13 in Epinions), (7 in ML-1m), (10 in ML-100k), and (11 in Ciao-i); conversely, RMSEs are the highest when d_s is set to the followings: (0 in FilmTrust, Ciao-u, Epinions, and Ciao-i), and (19 in ML-1m and ML-100k). The results show that it is important to balance the interaction and independence information rather than extreme consideration for one of them; note that more performance improvements are observed for larger and sparser data (Epinions, Ciao-u, and Ciao-i).

D. Effects of Data Context Adaptation (Q3)

We verify the effectiveness of data context adaptation layer in Figure 4. We compare three models: DAConA, DAConA-*withoutCA*, and DAConA-*sep*. DAConA is our proposed method where we optimize data context adaptation matrices D^X and D^Y in training process. DAConA-*withoutCA* does not adapt to each data context, by setting D^X and D^Y to fixed (non-learnable) identity matrices. To analyze the effect of setting a common adaptation matrix for each data context, we also devise DAConA-*sep* where user and item have separate data context adaptation matrices for the rating data context (e.g., D^{X-user} and D^{X-item} , respectively), and item and genre have separate data context adaptation matrices for the auxiliary data context (e.g., D^{Y-item} and $D^{Y-genre}$, respectively). For the three models, all hyperparameters are set to the optimal settings. Figure 4 shows RMSEs of the methods in the real world datasets. DAConA outperforms DAConA-*withoutCA* for all datasets, which means that context adaptation layer is effective for the performance. This is because DAConA projects the latent interaction vector through data context adaptation matrices that are trained to extract appropriate features for each data context, while DAConA-*withoutCA* reuses the same latent interaction vector in different data contexts; note

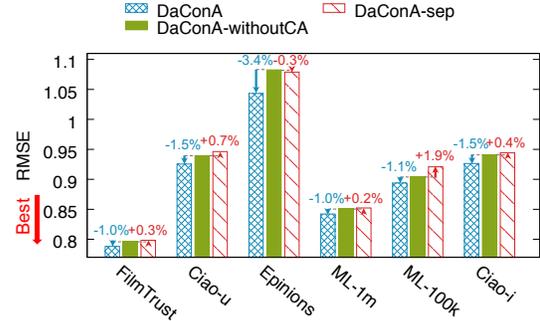


Fig. 4: DAConA outperforms DAConA-*withoutCA*, showing that data context adaptation improves the accuracy. DAConA also outperforms DAConA-*sep*, showing that sharing the same data context adaptation matrix improves the accuracy.

that more performance improvements are observed for larger and sparser data (Epinions, Ciao-u, and Ciao-i). Despite adding the learnable matrices, DAConA-*sep* shows little change in performance when compared to DAConA-*withoutCA*; rather, the performances are degraded in almost all datasets (except for Epinions). The results show that, in a data context, sharing the same data context adaptation matrix regardless of entities improves the generalization, while assigning different matrices per entity causes overfitting.

E. Neural Networks (Q4)

We investigate DAConA with different numbers of hidden layers to show whether deepening DAConA is beneficial to recommendation. We also present the effectiveness of activation functions by comparing versions of DAConA with/without activation functions in terms of RMSE.

First, we analyze whether deep layers of DAConA helps improve the performance in Figure 5. Since we set the depths of f_X and f_Y the same for all experiments, we define the depth of DAConA as the depth of f_X or that of f_Y . DAConA- k denotes DAConA with depth k . The structures of DAConA-1, DAConA-2, DAConA-3, DAConA-4, and DAConA-5 are [10], [20, 10], [40, 20, 10], [80, 40, 20, 10], and [160, 80, 40, 20, 10], respectively. Figure 5 shows RMSEs of DAConA-1 to DAConA-5 on real-world datasets. The results show that DAConA-3 has the best performance for all datasets. Although increasing the number of layers from 1 to 3 helps improve the performance thanks to increased model capacity, further increasing the number of layers hurts the performance possibly due to overfitting.

Second, we analyze whether non-linear activation functions of DAConA helps improve the performance in Figure 6. DAConA-*withoutAF* denotes DAConA without non-linear activation functions. The results show that the non-linear activation functions of DAConA improve the performance, which means activation functions are helpful in modeling non-linear relationships between entities in real-world datasets.

V. RELATED WORKS

Many studies proposed methods that leverage additional information to alleviate the rating sparsity problem in collabo-

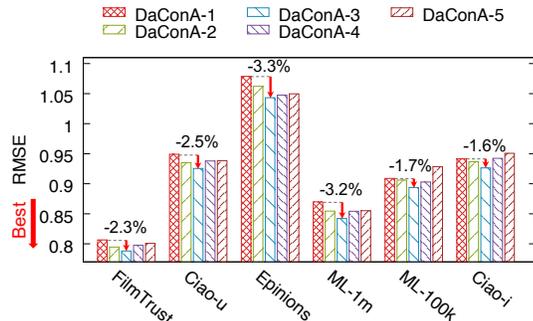


Fig. 5: Using 3 layers instead of too small or too large layers gives the best accuracy. DAConA- k denotes DAConA with k layers. Each percentage indicates the error reduction rate of the best model DAConA-3 compared to DAConA-1.

relative filtering. The methods are classified into two categories according to the type of additional data: user data context based methods, and item data context based methods.

User data context. Studies [12], [26]–[30] extensively proved that using not only rating data but also additional data about users mitigates rating sparsity problem in collaborative filtering. Ma et al. [12] proposed SoRec that uses rating matrix and truster-trustee matrix to predict unobserved ratings. SoRec jointly optimizes latent item vectors using both rating matrix and trust matrix. Yang et al. [29] proposed a method that extracts a subset of friends to be used in collaborative filtering. Guo et al. [27] proposed TrustSVD that combines user’s implicit data and rating matrix based on biased matrix factorization [1].

Item data context. Previous works [19]–[21], [31]–[34] proposed models that utilize item data context as additional information. Leung et al. [31] proposed a method that quantifies reviews through sentiment analysis and reflects it in rating prediction. Wang et al. [19] combined topic modeling and collaborative filtering. Wang et al. [20] integrated Stacked Denoising AutoEncoder (SDAE) [35] and Probabilistic Matrix Factorization (PMF) [2]. Dong et al. [11] combined two Additional Stacked Denoising AutoEncoders (aSDAE) to support side information of both user and item. Kim et al. [21] proposed ConvMF that integrates Convolutional Neural Networks capturing contextual information of item documents into the PMF. Hu et al. [32] proposed a model that integrates item reviews into Matrix Factorization based Bayesian personalized ranking (BPR-MF). Bauman et al. [33] proposed SULM that analyzes sentiment for each aspect by decomposing reviews into aspect units. SULM predicts not only the probability that a user likes an item, but also which aspect will have a large effect.

VI. CONCLUSION

We propose DAConA, a neural network based method that predicts missing rating values by exploiting rating matrix and auxiliary matrix. DAConA overcomes three challenges: modeling effective embedding vectors to represent *interaction* and *independence* information, considering data context difference, and modeling non-linear relationship between entities. We

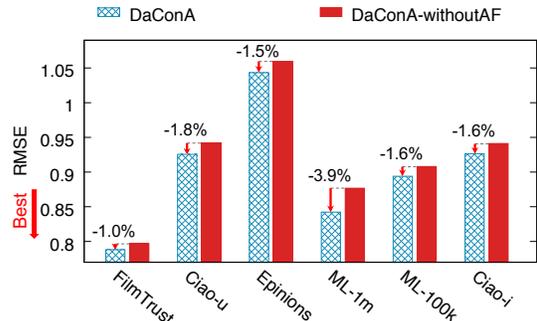


Fig. 6: Non-linear activation function helps improve the accuracy in DAConA.

separate the embedding vectors into latent interaction vectors and rich latent independence vectors for entities. DAConA uses data context adaptation layer to extract latent factors suitable for each data context. Furthermore, DAConA uses neural networks to model complicated non-linear relationship between latent factors. We show that DAConA is a generalized method which integrates many methods including MF, biased-MF, CMF, and biased-CMF as special cases. Through extensive experiments we show that DAConA gives the state-of-the-art performance in recommendation with rating data and auxiliary matrix. Future works include extending the method for a time-evolving setting.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) funded by MSIT(2019R1A2C2004990). U Kang is the corresponding author.

REFERENCES

- [1] Y. Koren, R. M. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *IEEE Computer*, 2009.
- [2] R. Salakhutdinov and A. Mnih, “Probabilistic matrix factorization,” in *NIPS*, 2007.
- [3] H. Park, J. Jung, and U. Kang, “A comparative study of matrix factorization and random walk with restart in recommender systems,” in *IEEE BigData*, 2017.
- [4] G. K. Dziugaite and D. M. Roy, “Neural network matrix factorization,” *CoRR*, 2015.
- [5] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, “Neural collaborative filtering,” in *WWW*, 2017.
- [6] A. P. Singh and G. J. Gordon, “Relational learning via collective matrix factorization,” in *SIGKDD*, 2008.
- [7] S. Rendle, “Factorization machines,” in *ICDM*, 2010.
- [8] S. Rendle, “Factorization machines with libfm,” *ACM TIST*, 2012.
- [9] M. Blondel, A. Fujino, N. Ueda, and M. Ishihata, “Higher-order factorization machines,” in *NIPS*, 2016.
- [10] C. Liu, C. Zhou, J. Wu, Y. Hu, and L. Guo, “Social recommendation with an essential preference space,” in *AAAI*, 2018.
- [11] X. Dong, L. Yu, Z. Wu, Y. Sun, L. Yuan, and F. Zhang, “A hybrid collaborative filtering model with deep structure for recommender systems,” in *AAAI*, 2017.
- [12] H. Ma, H. Yang, M. R. Lyu, and I. King, “Sorec: social recommendation using probabilistic matrix factorization,” in *CIKM*, 2008.
- [13] H. Ma, I. King, and M. R. Lyu, “Learning to recommend with social trust ensemble,” in *SIGIR*, 2009.
- [14] M. Jamali and M. Ester, “A matrix factorization technique with trust propagation for recommendation in social networks,” in *RecSys*, 2010.
- [15] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, “Recommender systems with social regularization,” in *WSDM*, 2011.
- [16] B. Yang, Y. Lei, D. Liu, and J. Liu, “Social collaborative filtering by trust,” in *IJCAI*, 2013.

- [17] J. Tang, S. Wang, X. Hu, D. Yin, Y. Bi, Y. Chang, and H. Liu, "Recommendation with social dimensions," in *AAAI*, 2016.
- [18] S. Li, J. Kawale, and Y. Fu, "Deep collaborative filtering via marginalized denoising auto-encoder," in *CIKM*, 2015.
- [19] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *SIGKDD*, 2011.
- [20] H. Wang, N. Wang, and D. Yeung, "Collaborative deep learning for recommender systems," in *SIGKDD*, 2015.
- [21] D. H. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *RecSys*, 2016.
- [22] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artificial Intelligence*, 2009.
- [23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [25] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *RecSys*, 2016.
- [26] A. J. Chaney, D. M. Blei, and T. Eliassi-Rad, "A probabilistic model for using social networks in personalized item recommendation," in *RecSys*, 2015.
- [27] G. Guo, J. Zhang, and N. Yorke-Smith, "Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings," in *AAAI*, 2015.
- [28] G. Guo, J. Zhang, D. Thalmann, and N. Yorke-Smith, "ETAF: an extended trust antecedents framework for trust prediction," in *ASONAM*, 2014.
- [29] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks," in *SIGKDD*, 2012.
- [30] H. Ma, "An experimental study on implicit social recommendation," in *SIGIR*, 2013.
- [31] C. W. Leung, S. C. Chan, and F.-I. Chung, "Integrating collaborative filtering and sentiment analysis: A rating inference approach," in *ECAI workshop on recommender systems*, 2006.
- [32] G. Hu and X. Dai, "Integrating reviews into personalized ranking for cold start recommendation," in *PAKDD*, 2017.
- [33] K. Bauman, B. Liu, and A. Tuzhilin, "Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews," in *SIGKDD*, 2017.
- [34] L. Chen, G. Chen, and F. Wang, "Recommender systems based on user reviews: the state of the art," *User Model. User-Adapt. Interact.*, 2015.
- [35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *JMLR*, 2010.

APPENDIX

A. Using Multiple Auxiliary Information

Extension. DAConA supports using multiple auxiliary matrices. Assume we are given an items' auxiliary matrix and a users' auxiliary matrix. Let $X \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ be a rating matrix, $Y \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{C}|}$ be an items' auxiliary matrix, and $Z \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{T}|}$ be a users' auxiliary matrix. In the following, let Y denote a movie-genre matrix, and Z denote a user-trustee matrix. Then \mathcal{U} , \mathcal{I} , \mathcal{C} , and \mathcal{T} indicate sets of users, items, genres, and trustees, respectively. Then the objective function is extended as follows:

$$L = (1 - \alpha - \beta)loss_X + \alpha loss_Y + \beta loss_Z, \quad (9)$$

where $loss_X$ is Equation (2), $loss_Y$ is Equation (3), and

$$loss_Z = \frac{1}{2} \sum_{(i,l) \in \Omega_Z} (\hat{Z}_{il} - Z_{il})^2 + \frac{\lambda}{2} Reg_Z. \quad (10)$$

Ω_Z contains observable entries in Z , \hat{Z}_{il} is predicted entry, and Reg_Z is a regularization term. The predictive model for Z is defined as follows:

$$\hat{Z}_{il} = f_Z \left(\begin{bmatrix} D^Z U_i \circ D^Z T_l \\ U_i^Z \\ T_l^Z \end{bmatrix} \right), \quad (11)$$

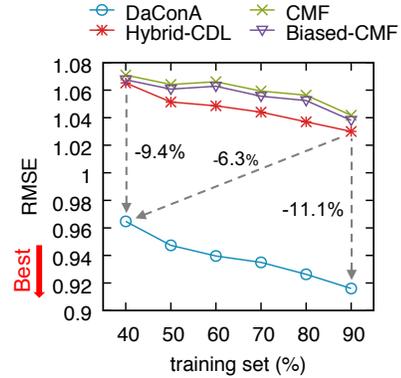


Fig. 7: DAConA gives the best accuracy even when using two auxiliary matrices. Each arrow indicates the reduction rate of DAConA's RMSE compared to hybrid-CDL's RMSE.

where f_Z is a fully-connected neural network. The predictive models for X and Y are defined in Equation (6) (details in Section III-E). T is the latent interaction matrix for the trustees, U^Z is the latent independence matrix for the users in the user-trustee data context, and T^Z is the latent independence matrix for the trustees in the user-trustee data context; note that the latent interaction matrix U of users is trained from both X and Z ; similarly, V is trained from both X and Y . Using this approach, it is trivial to further extend DAConA to utilize more than two auxiliary matrices (e.g., movie-director, movie-document, and user-demographic).

Experiments. Among the competitors, CMF [6], Biased-CMF, and Hybrid-CDL [11] which use multiple auxiliary information are chosen for detailed comparison; even though FM also utilizes multiple auxiliary information, its poor performance clearly removes the need for detailed comparison (see Table IV). We compare the models in Ciao² dataset which consists of a rating matrix and two auxiliary matrices: item-genre matrix and user-trustee matrix. CMF and Biased-CMF are composed of three MFs and three Biased-MFs, respectively. As described in Section IV-A, Hybrid-CDL uses two Additional Stacked Denoising Autoencoders (aSDAE), one for users and the other for items. We find the optimal hyperparameters by grid search. For CMF and Biased-CMF, we set l and λ as $1e-4$ and $1e-4$, respectively. For hybrid-CDL, we adopt [500, 10, 500]-structured hidden layers to the two aSDAEs, and set l , λ , α_1 , and α_2 as $1e-3$, $1e-3$, 0.8, and 0.3, respectively; we also use corrupted inputs with a noise level of 0.3 as described in [11]. For DAConA, we use [40, 20, 10]-structured hidden layers, and set l , λ , d_s , α , and β as $1e-3$, $1e-5$, 11, 0.6, and 0.2, respectively. Figure 7 shows the RMSEs while varying the percentage of training set; the rest of the data are used as test set. Note that DAConA consistently outperforms the competitors by a significant margin, better utilizing auxiliary information. Even DAConA using 40% of training data shows 6.3% lower RMSE than hybrid-CDL using 90% of training data.