Collaborative Filtering with Generalized Laplacian Constraint via Overlapping Decomposition

Qing Zhang, Houfeng Wang
Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, China
{zqicl,wanghf}@pku.edu.cn

Abstract

Real-world data are seldom unstructured, yet traditional Matrix Factorization (MF) models, as one of the most powerful collaborative filtering approaches, generally rely on this assumption to recover the low-rank structures for recommendation. However, few of them are able to explicitly consider structured constraint with the underlying low-rank assumption to model complex user interests. To solve this problem, we propose a unified MF framework with generalized Laplacian constraint for collaborative filtering. We investigate the connection between the recently proposed Laplacian constraint and the classical normalized cut problem, and make it possible to extend the original non-overlapping prior, to capture the overlapping case via learning the decomposed multi-facet graphs. Experiments on real-world datasets demonstrate the effectiveness of the proposed method.

1 Introduction

Collaborative Filtering (CF) algorithms [Salakhutdinov and Mnih, 2007] have been widely applied in various recommender systems. As one of the most powerful CF approaches, Matrix Factorization (MF) models have become popular and achieve the state-of-the-art performance [Zhang *et al.*, 2013]. The rationale behind the MF approach is the low-rank assumption, that the observed rating data can be explained holistically by a small number of latent factors for both users and items, which, for instance, may be related to user groups and item topics. However, traditional MF models rarely consider the structured constraint explicitly on this underlying low-rank structure. In previous studies, incorporating structured information has been extensively explored based on side information, such as social network [Purushotham and Liu, 2012] and item content [Wang and Blei, 2011].

However, not much research work has been done towards directly incorporating the constraint from optimization perspective in a fundamental setting [Yuan *et al.*, 2014]. The reason is that it is hard to optimize with structured constraint during learning, and it is hard to pre-define a fixed structure as constraint to capture the complex latent user interests. In real applications, one user may have multiple interests, which is usually characterized by multiple views [Gao *et al.*, 2013]. For example, clustering users based on music reviews can be clustered by a genre (rock, jazz, hip hop, etc.) or a sentiment (positive, negative, etc.). This may lead to different solutions for modeling user hidden structures. Motivated by this observation, we generalize the recently proposed non-overlapping Laplacian constraint [Feng *et al.*, 2014] to the overlapping case, with the ability to automatically capture such multi-view hidden structures for collaborative filtering. The mechanism proposed in this paper is significantly different from the emerging group-constraint CF methods as shown in Table 1, since this work is proposed from an adaptive multi-view perspective in a joint optimization framework, without the non-overlapping assumption [Zhang and Wang, 2015] or pre-partition restriction [Yuan *et al.*, 2014] for hidden group (community) structures.

Method	Adap.	A.G.	G.O.	Constraint
GSMF	$\surd$	$\times$	$\surd$	L_{21} norm (factor)
LMF	$\times$	$\surd$	$\surd$	permutation
HGMF	$\surd$	$\surd$	$\times$	Laplacian
This work	$\surd$	$\surd$	$\surd$	L_{21} norm (view)+HSIC

Table 1: Summary of this work and other related methods with group-based constraint. Adap.: Adaptive to CF objective. A.G.: Automatic Grouping. G.O.: Group Overlapping. GSMF [Yuan *et al.*, 2014]; LMF [Zhang *et al.*, 2013]; HGMF [Zhang and Wang, 2015].

2 Problem Formulation and Preliminaries

Definition 1 (Matrix Factorization Models)**.** *Given a sparse rating matrix $R = [r_{ij}]$, where the observed r_{ij} denotes the rating of user i on item j, matrix factorization models aim to factorize $R = UV^T$, where U and V are low rank matrices with rows as latent users u_i^T and latent items v_j^T respectively.*

Our goal is to predict the missing values in R, by computing the predicted values $r_{ij} = u_i^T v_j$.

Laplacian Constraint for Block-diagonality

We introduce the recently proposed structured constraint, i.e., Laplacian constraint [Feng *et al.*, 2014], to recommendation tasks, for explicitly capturing the latent user community

structures while learning matrix factorization. We first define Laplacian matrix, and then present its connection with the structure of affinity matrix.

Definition 2 (Laplacian Matirx)**.** *Consider an affinity matrix $W \in \mathbb{R}^{n\times n}$ of n samples with weights $W(i,i')$. The Laplacian matrix $L_W \in \mathbb{R}^{n\times n}$ is defined as: $L_W = D - W$, where $D = diag(d_1, ..., d_n)$ and $d_n = \Sigma_{i'} W(i,i')$. The normalized version is defined as $L_{W_{sys}} = D^{-\frac{1}{2}} L_W D^{-\frac{1}{2}}$.*

The following well known theorem relates the rank of the Laplacian matrix to the number of blocks in W.

Theorem 1 ([von Luxburg, 2007])**.** *Let W be an affinity matrix. The multiplicity k of the eigenvalue 0 of the Laplacian $L_{W_{sys}}$, equals the number of connected blocks in W.*

Based on the above theorem, we can enforce a general square matrix to be *k*-block-diagonal to represent different latent communities. For the hidden graph constructed by latent users u_j, we construct the affinity matrix $W(j,j')$ using Gaussian kernel. Then we can define a set of k-block-diagonal matrix (*k*-BDMS) as the constraint term in Eq.(2),

$$\begin{aligned}\mathcal{K} =& \{W | rank(L_{W_{sys}}) = n - k, \\ & W(i,i') = w_{ii'} = exp(-\frac{\|u_i - u_{i'}\|_2^2}{\sigma^2})\},\end{aligned} \tag{1}$$

where $\|.\|_2^2$ denotes the $\ell 2$ norm, and σ^2 denotes the deviation.

Adaptive Hidden Graph Regularization Framework

To incorporate the above structured constraint, we employ an adaptive hidden graph regularization framework [Zhang and Wang, 2015], which is similar to [Purushotham and Liu, 2012], but the graph used for regularization is learnt automatically with structured prior $W \in \mathcal{K}$, rather than the pre-defined one based on side information. To achieve the goal, we have the following optimization objective:

$$\begin{aligned}\min_{U,V,W,S} & \frac{1}{2}\sum_{i,j} c_{ij}(r_{ij} - u_i v_j^T)^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2 \\ & + \frac{\lambda_W}{2}\sum_{i,i'} c_{w_{ii'}}(w_{ii'} - u_i s_{i'}^T)^2 + \frac{\lambda_S}{2}\|S\|_F^2, \\ & s.t.\ W \in \mathcal{K}\end{aligned} \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm; W is the same one in Eq.(1); c_{ij} is a confidence parameter [Wang and Blei, 2011] for rating r_{ij}, and $c_{w_{ii'}}$ is similarly defined for modeling $w_{ii'} \in W$; $\lambda_{U,V,S}$ are used to avoid over-fitting; λ_W controls the effect of hidden graph regularization; The first term ensures that latent users U and items V can well approximate the observed ratings $r_{ij} \in R$. Similarly, the fourth term factorizes W into a user-specific matrix U and a factor-specific matrix $s_{i'} \in S$, to approximate $w_{ii'}$ with constraint $W \in \mathcal{K}$.

Previous Non-Overlapping Solution [Feng *et al.*, 2014; Zhang and Wang, 2015]. To learn $W \in \mathcal{K}$, after updating U, V, S, we can obtain the hidden graph W_0 using Gaussian kernel in Eq.(1). However, the variable matrix W_0 may move out of the constraint set and no longer satisfy a k-block-diagonal structure. To project it back to the k-BDMS constraint set, we have the following optimization problem via Augmented Lagrangian Multiplier method [Lin *et al.*, 2011]:

$$\begin{aligned}\min_{W,\tilde{Z}} & \frac{1}{2}\|W - W_0\|_F^2 + \langle J, \tilde{Z} - L_{W_{sys}}\rangle + \\ & \frac{\beta}{2}\|\tilde{Z} - L_{W_{sys}}\|_F^2, \quad s.t.\ rank(\tilde{Z}) = n - k\end{aligned} \tag{3}$$

where J is the Lagrangian multiplier and β is an increasing weight parameter for the term of enforcing the auxiliary variable $\tilde{Z} = L_{W_{sys}}$.

3 Generalizing Laplacian Constraint

Although the Laplacian constraint could well capture hidden group structures, the non-overlapping assumption behind it may be too strong in real-world scenarios. In real world, data can often be interpreted in many different ways, which can have different groupings that are reasonable and interesting from different perspectives. In this section, we propose a solution to generalize Laplacian constraint, which relaxes the non-overlapping assumption for real applications.

Connection between Laplacian Constraint and Normalized Cut

For our extension, we first reveal the connection between the recently proposed Laplacian constraint as shown in Eq.(3) and Normalized Cut problem as defined in Def.(3) from optimization perspective.

Definition 3 (Normalized Cut (Ncut))**.** *Given a similarity graph with affinity matrix W, for two disjoint subsets A, $B \in V$, we define $cut(A,B) = \sum_{i\in A, j\in B} w_{ij}$. For arbitrary k disjoint subsets $A_1, ..., A_k$, we define $cut(A_1, ..., A_k) = \sum_{i=1}^k cut(A_i, \overline{A}_i)$. The normalized cut (Ncut) is defined as:*

$$Ncut(A_1, ..., A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{vol(A_i)}, \tag{4}$$

where $vol(A_i) = \sum_{i\in\in A_i, i'\in A_i} w_{ii'}$, the total weights of the edges in group A_i.

The intuition of minimizing Ncut [von Luxburg, 2007] is to find a partition of the graph, such that the edges between different groups have low similarity, and those within a group have high similarity.

Theorem 2 (Rank Constraint Approximation)**.** *The solution to Eq.(3) can be approximated in terms of rank constraint, through solving minimal Ncut problem, with the original hidden affinity matrix W_0 in Eq.(3).*

Proof. As shown in [von Luxburg, 2007; Shi and Malik, 2000], solving minimal Ncut problem is equivalent to solving the following trace minimizing problem: $\min_{U\in R^{n\times k}} Tr(U^T D^{-\frac{1}{2}} L_W D^{-\frac{1}{2}} U) = \sum_{i=1}^k \lambda_i$, with constraint $U^T U = I$, where λ_i is the eigenvalue of matrix $U^T D^{-\frac{1}{2}} L_{W_{sys}} D^{-\frac{1}{2}} U$. Note $L_{W_{sys}} = D^{-\frac{1}{2}} L_W D^{-\frac{1}{2}}$. Based on this trace minimizing problem, our goal is to prove λ_i also to be the ith smallest eigenvalue of matrix $L_{W_{sys}}$. According to Courant-Fisher theorem [Golub and van Loan, 1996], the optimal U contains the first smallest k eigenvectors of matrix $L_{W_{sys}}$ as columns. Now, use the optimal U.

$U^T L_{W_{sys}} U$ can be written as $U^T L_{W_{sys}} U = U^T L_{W_{sys}}[U(:,1),...,U(:,k)] = U^T[L_{W_{sys}}U(:,1),...,L_{W_{sys}}U(:,k)] = U^T[\lambda_1 U(:,1),...,\lambda_k U(:,k)] = [\lambda_1 U^T U(:,1),...,\lambda_k U^T U(:,k)] = [\lambda_1 e_1,...,\lambda_k e_k] = diag(\lambda_1,...,\lambda_k)$, where e_k is a column vector containing 1 in kth position and zeros for others. Then, $tr(U^T L_{W_{sys}} U) = tr(diag(\lambda_1,...,\lambda_k)) = \sum_{i=1}^{k} \lambda_i$., which shows that minimizing Ncut problem is equivalent to minimizing the top-k smallest eigenvalues of the Laplacian $L_{W_{sys}}$. Finally, since the rank of a positive semi-definite (PSD) matrix is equal to the number of nonzero eigenvalues, the rank of $L_{W_{sys}}$ will approximate $n-k$ to satisfy the constraint in Eq.(3), from which the theorem follows. □

Theorem 3 (Upper Error Bound)**.** *Let W and W_0 be an approximatively projected and original affinity matrix respectively. Then in the sense of approximately minimizing the first smallest k eigenvalues of the Laplacian matrix $L_{W_{sys}}$, the error between W and W_0 is upper bounded by*

$$\|W - W_0\|_F \leq \|W_0\|_F \|Q'Q'^T W_0 - I\|_F, \qquad (5)$$

where $Q' = [\frac{Q_0 x_{\lambda_1}}{\lambda_1}, \ldots, \frac{Q_0 x_{\lambda_k}}{\lambda_k}]$. x_{λ_k} is a normalized eigenvector of $W_0 Q_0$ with eigenvalue λ_k. $Q_0 = U_0 V_0^T$. U_0 and V_0 are the solutions of computing SVD on $W_0^T L_{W_{sys}}$.

Proof. We first seek to represent W by a certain form involving W_0. Suppose $L_{W_{sys}}$ is the normalized Laplacian matrix of W_0. The optimal solution of W is determined by the first smallest k eigenvectors of $L_{W_{sys}}$ for spectral clustering. To establish the relationship between W and W_0, we formulate the following optimization problem, $\min_{Q_0} \|L_{W_{sys}} - W_0 Q_0\|_F$, $s.t. Q_0^T Q_0 = I$. This problem can be seen as a orthogonal Procrustes problem [Golub and van Loan, 1996], which has the optimal solution $Q_0 = U_0 V_0^T$, where U_0 and V_0 are the solutions of computing SVD on $W_0^T L_{W_{sys}}$. Since $L_{W_{sys}} v_i = v_i \lambda_i$ and $V = [v_1,...,v_k]$ can be seen as the latent feature matrix, then we can approximatively construct W using V to obtain $W = V * V^T$. Using the solution of the above formulated optimization problem, we can derive $v_k = \frac{L_{W_{sys}} v_k}{\lambda_k} \approx \frac{W_0 Q_0 v_k}{\lambda_k}$, by substituting $L_{W_{sys}}$ with the optimal $W_0 Q_0$. Let $Q' = [\frac{Q_0 x_{\lambda_1}}{\lambda_1}, \ldots, \frac{Q_0 x_{\lambda_k}}{\lambda_k}]$. Now we can easily reformulate, $\|W - W_0\|_F \approx \|W_0 Q' Q'^T W_0 - W_0\|_F = \|W_0(Q'Q'^T W_0 - I)\|_F \leq \|W_0\|_F \|Q'Q'^T W_0 - I\|_F$, from which the theorem follows. □

We have proven that solving Ncut problem as spectral clustering objective, can approximate the original problem in Eq.(3), because it can achieve the same optimization objective, in terms of approximating the original rank constraint (Theorem 2), with bounded graph projection error (Theorem 3).

Overlapping Decomposition via Non-overlapping Multi-facet Graphs

Overlapping is a significant property of real-world structures. However, the original Laplacian constraint could not capture this property. *Motivated* by the established connection between Laplacian constraint and spectral clustering, we can extend it to handle the overlapping case from the graph clustering view. *Without changing the basic framework* in Eq.(2), we develop a novel structured prior, as constraint $s.t.\ W \in \mathcal{K}'$ in Eq.(2) for collaborative filtering. The key idea is to decompose a complex real-wold overlapping case, into a series of non-overlapping cases ($\mathcal{K}'$), as shown in Figure 1.

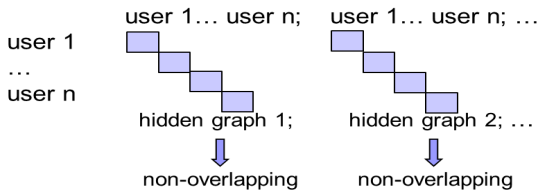


Figure 1: Modeling Hidden Overlapping Communities via Multi-facet Graphs. It shows one generalized graph for regularization, by concatenating two multi-facet graphs directly. If we permutate the rows to align the two non-overlapping graphs, by satisfying the same user appearing in the same row, the concatenated graph could be overlapping.

Definition 4 (Multi-facet Graphs)**.** *Consider an affinity matrix $W^i \in \mathbb{R}^{n \times n}$ of n samples with weights $W^i(i,i')$. The Multi-facet Graphs $W^1,...,W^i$ are defined as: each W^i is a non-overlapping graph with block-diagonality. For different graphs, W^i and W^j where $i \neq j$, are expected to reflect the different views of the data.*

Multi-facet Graphs. We define multi-facet graphs in Def.(4), using multiple non-overlapping graphs with different views to capture the original overlapping information. Ideally, these decomposed non-overlapping graphs are expected to reflect different views of the original graph. We propose to achieve the goal, through automatic latent factor selection to construct different views while learning matrix factorization.

3.1 Algorithm

Algorithm 1: Algorithm for Learning Multi-facet Graphs

Input: Number of views Q, latent user embedding matrix U, λ_1 for HSIC penalty, λ_2 for group sparsity.
Output: $U_q, W_q, q = 1,2,...,Q$.

```
while q < Q do
    Update each W_q according to Section 3.1;
    Update each U_q according to Section 3.1;
    q = q + 1;
end while
Return U_q, W_q, q = 1,2,...,Q;
```

To acquire such multi-facet graphs, inspired by the research of multi-facet clustering [Niu *et al.*, 2010], which believes that document is often composed of multiple views. *In this paper, we extend this idea to the hidden graph setting jointly with feature selection* [Masaeli *et al.*, 2010].

We first define the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005] between two random variables (X,Y). Given n observations, $Z := \{(x_1,y_1),...,(x_n,y_n)\}$, HSIC can be empirically estimated by $HSIC(X,Y) = (n-$

$1)^{-2}tr(K_1HK_2H)$, where $K_1, K_2 \in R^{n\times n}$ are Kernel matrices, $(K_1)_{ij} = k_1(x_i, x_j)$, $(K_2)_{ij} = k_2(y_i, y_j)$, and where $(H)_{ij} = \delta_{ij} - n^{-1}$ centers the Kernel matrices to have zero mean in the feature space, where δ_{ij} is Kronecker delta. To achieve learning multi-view hidden structures, we use the HSIC as a penalty term in our spectral clustering objective function to ensure that subspaces in different hidden graphs provide non-redundant information. Moreover, to achieve the goal of automatic feature selection for each view, we propose to do feature selection and hidden graph learning simultaneously in a joint framework, which is based on l_{21}-norm that is used on the projection matrix W_q.

$$\begin{aligned}\min_{W_q,U_q} &\sum_q Tr(U_q^T D_q^{-\frac{1}{2}} L_W^q D_q^{-\frac{1}{2}} U_q) \\ &+ \lambda_1 \sum_{q\neq r} HSIC(W_q^T x, W_r^T x) + \lambda_2 \sum_q \|W_q\|_{2,1} \\ s.t.\ &U_q^T U_q = I,\ W_q^T W_q = I\end{aligned} \tag{6}$$

Interpretation. In the above equation, the first term with orthogonal constraint is the spectral clustering objective, to learn hidden group structures in different views. The second term with orthogonal constraint is the multi-facet penalty, to maximize the discrepancy between different views. The third term is the group sparsity penalty, to automatically select latent factors for learning hidden group structures in different views. $x = U^T$ is the latent embeddings as column vectors for representing each user in original latent factor space. Note that the notation U_q is the spectral embeddings, while U is the latent user embeddings. L_W^q is constructed by gaussian kernel with the projected data in each view $W_q^T x$. In the following, we use x_i to represent the *i*-th latent user representation.

Discussion on $\|W_q\|_{2,1}$. The $l_{2,1}$-norm of a matrix was first introduced in [Ding *et al.*, 2006] as rotational invariant $l_{2,1}$-norm and also used as group sparsity penalty for subspace learning [Gu *et al.*, 2011]. It is defined as $\|W_q\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m (w_{ij}^q)^2} = \sum_{i=1}^n \|\mathbf{w}_q^{i\cdot}\|_2$, where $\mathbf{w}_q^{i\cdot}$ is the ith row vector. To understand why the proposed $\|W_q\|_{2,1}$ penalty leads to a feature selection solution, let us consider how the structure of W_q should be to achieve feature selection. Let w_{ij}^q be the elements of transformation matrix W_q. If ith feature in x is not selected by the hidden graph in view q, all the elements of the ith row of W_q should be zero. Thus, ith feature in x will not contribute to the $HSIC$ criterion.

Optimize for W_q for each view. Fixing U_q, optimizing W_q is *challenging*, since this optimization problem is generally difficult to optimize due to the non-convexity of orthogonal constraints. To solve this challenge, we employ an efficient feasible method for optimization on the Stiefel manifold [Bach and Jordan, 2002]. Following the method, we apply a *modified gradient descent* on the Stiefel manifold, to ensure the orthogonal constraints to be preserved in each iteration:

Step1: We project the negative gradient of the objective function onto the tangent space [Trendafilov, 2010], $\Delta W_{Stiefel} = -\frac{\partial f}{\partial W_q} - W_q(-\frac{\partial f}{\partial W_q})^T W_q$. **Step2**: We update W_q on the geodesic [Trendafilov, 2010] in the direction of the tangent space. $\Delta W_{new} = W_{old} exp(\tau W_{old}^T \Delta W_{Stiefel})$, where exp means matrix exponential and τ is the step size. **Step3**: We use the Armijo rule for a backtracking line search to find the step size τ at every iteration.

More specifically, the derivative $\frac{\partial f}{\partial W_q}$ is calculated as follows. There are three terms involved for computing the partial derivative with respect to W_q, in the objective function for each view. Note that $D_q^{-\frac{1}{2}} L_W^q D_q^{-\frac{1}{2}} = I - D_q^{-\frac{1}{2}} K_q D_q^{-\frac{1}{2}}$.

i) The spectral objective term: We define $L^q = D_q^{-\frac{1}{2}} K_q D_q^{-\frac{1}{2}}$. The spectral objective can be expressed as a linear combination of each element in matrix L^q with the corresponding element in $U_q U_q^T$ as coefficient. The derivative of the element l_{ij}^q in L^q is, $(l_{ij}^q)' = (k_{ij}^q)'(d_{ii}^q)^{-\frac{1}{2}}(d_{jj}^q)^{-\frac{1}{2}} - \frac{1}{2}(d_{ii}^q)^{-\frac{1}{2}}(d_{ii}^q)' k_{ii}^q (d_{jj}^q)^{-\frac{1}{2}} - \frac{1}{2}(d_{jj}^q)^{-\frac{1}{2}}(d_{jj}^q)' k_{ij}^q (d_{ii}^q)^{-\frac{1}{2}}$, where $(k_{ij}^q)'$, $(d_{ii}^q)'$, $(d_{jj}^q)'$ are derivatives of the similarity degree matrix with respect to W_q.

ii) The empirical HSIC term: We expand the trace in the HSIC term, $tr(K_q H K_r H) = tr(K_q K_r) - 2n^{-1}\mathbf{1}^T K_q K_r \mathbf{1} + n^{-2} tr(K_q) tr(K_r)$, where $\mathbf{1}$ is the vector of all ones. Using a Gaussian kernel defined as $k(W_q^T x_i, W_q^T x_j) = exp(-\|W_q^T \Delta x_{ij}\|^2/2\sigma^2)$, where Δx_{ij} is $x_i - x_j$, the derivative of k_{ij}^q in K_q is, $\frac{\partial k_{ij}^q}{\partial W_q} = -\frac{1}{\sigma^2}\Delta x_{ij} \Delta x_{ij}^T W_q exp(\frac{-\Delta x_{ij}^T W_q W_q^T \Delta x_{ij}}{2\sigma^2})$.

iii) The group sparsity term: The derivative of the column vector in $\|W_q\|_{2,1}$ is, $\frac{\partial \|W_q\|_{2,1}}{\partial W_q^{\cdot j}} = R_q W_q^{\cdot j} = diag(..., \frac{1}{\|W_q^{i\cdot}\|_2}, ...) W_q^{\cdot j}$, where R_q is a diagonal matrix, $W_q^{\cdot j}$ and $W_q^{i\cdot}$ are the jth column and ith row vectors respectively. Note that R_q is dependent to W_q, we propose to update R_q iteratively too, using the efficient strategy in [Nie *et al.*, 2010]. In each iteration, R_q is calculated with the current W_q, and then W_q is updated based on the current calculated R_q.

Optimize for U_q in each view. Fixing W_q, we can optimize U_q. The solution for U_q is equal to the first c_q eigenvectors (corresponding to the smallest c_q eigenvalues) of the matrix $D_q^{-\frac{1}{2}} L_W^q D_q^{-\frac{1}{2}}$, where c_q is the number of clusters for view q. Then we normalize each row of U_q to have unit length [Ng *et al.*, 2001; Kumar and III, 2011].

Construct W via Multi-facet Graphs. After updating U, V, S (the update rules are omitted, which is similar to [Zhang and Wang, 2015]) and running **Algorithm 1**, we can update a new hidden graph W as adaptive graph regularization term, for the *joint optimization problem* in Eq.(2). We construct W via multi-facet graphs, by directly concatenating $W = [W'_{q=1}, ..., W'_{q=Q}]$. *For pursuing the sparsity of W, to save computing and enlarge discriminative power*, we adopt a discretization way to build the graph W'_q in view q. In this way, we use simple k-means algorithm to partition the data using each row in U_q as features, and label each data according the group assignment. Then for the data in the same group, the weights between those are set to 1, or 0 otherwise.

Computational Complexity. For large-scale data, to calculate kernel matrix, we can apply incomplete Cholesky decomposition as suggested in [Bach and Jordan, 2002], in which the complexity of eigen-decomposition is $O(s^2 n)$ ($s \ll n$), where s is the rank of the approximation matrix, n is the number of general users. Therefore, the main com-

Data	#User	#Item	Sparsity	#Rating	#Avg.R
LastFM	1892	18745	0.28%	92834	49
Delicious	1867	69223	0.08%	104799	56
Epinions	3474	26850	0.12%	111933	32

Table 2: R denotes #rated per user. The rating in LastFM and Delicious is binary, and that in Epinions is a value [1,5].

putational cost of the algorithm is to calculate the SVD for U_q, which can be efficiently solved in $O(nmd)$ $(m, d \ll n)$ [Li *et al.*, 2011], where m is the number of selected subset of data; d is number of latent factors. Thus, similar to [Niu *et al.*, 2010], since the complexities of our derivative computation is $O(nsd)$ $(s, d \ll n)$, Algorithm 1 can achieve the linear complexity $O(s^2n + nmd + nsd)$ with respect to general users.

4 Experiments

We evaluate our method on three public real-world datasets as shown in Table 2: **LastFM** and **Delicious**[1] are used for Top-N recommendation task. **Epinions**[2] is used for rating prediction task. We use Recall [Purushotham and Liu, 2012] metric to measure Top-N performance, and use RMSE [Yuan *et al.*, 2014] metric to measure rating prediction performance.

Baselines and Settings

We compare the following popular and the state-of-the-art constraint MF models, only using rating information. **WNMF** [Zhang *et al.*, 2006] is a popular MF method, using non-negative matrix factorization. **PMF** [Salakhutdinov and Mnih, 2007] is a well-known MF method, with sound probabilistic interpretation. **GSMF** [Yuan *et al.*, 2014] is a constraint MF method, using group sparsity regularization for modeling multiple user interests. **GSMF-K** denotes using k-means to partition the item set into K groups for GSMF. **HGMF** [Zhang and Wang, 2015] is a constraint MF method, in an adaptive graph regularization framework using Laplacian constraint. **MHGMF** is the proposed method in this paper.

Following [Purushotham and Liu, 2012; Yuan *et al.*, 2014], we randomly select $90(70, 50)\%$ of the data for training, and the rest for testing. The random selection was carried out 5 times independently, and we report the average results. Using the widely adopted strategy in [Hu *et al.*, 2008], all the baselines can be applied to the both tasks. We implement the compared methods following the original works. The best parameters are chosen by held-out validation. For our method, the kernel parameter σ^2 is initialized as the median pairwise distance between original users, and then is well tuned. For top-N task, the confidence parameters are 0.01 and 1 for unobserved data and observed data respectively. For rating prediction task, those are 0 and 1 respectively. The number of multi-facet graphs is 2. The regularization parameters are optimized in the range of $\{0.001, 0.01, ..., 10, 100\}$. The numbers of latent factors, for the original space and the projected space, are set by searching $\{50, 150, ..., 350, 400\}$. We set the number of maximal iteration to 200.

[1]http://grouplens.org/datasets/hetrec-2011/

[2]http://www.public.asu.edu/∼jtang20/datasetcode/epinions.zip

Results and Analysis

Performance Comparison. Table 3 and Figure 2(a,b) show that in general, the constraint based models, i.e., MHGMF; HGMF; GSMF-K, are much better than the non-constraint models, i.e., WNMF; PMF, which demonstrates the effectiveness of the underlying structured assumption in real-world data. In addition, the results show that our method MHGMF further outperform the existing constraint based models on all evaluation metrics on all three datasets, which could demonstrate that it is more reasonable to use the proposed overlapping assumption for real-world data from multiple views, compared with the non-overlapping assumption in HGMF from a single view. Although GSMF can deal with overlapping group structures, it is limited to the need of the pre-partitioned groups, without adaptive grouping mechanism for the task objective. In contrast, our method can consider the structured property jointly with task-oriented objective from an adaptive multi-view setting. In the following, we show the parameter sensitivity analysis with 90% training case, due to the similar findings for other cases and page limitation.

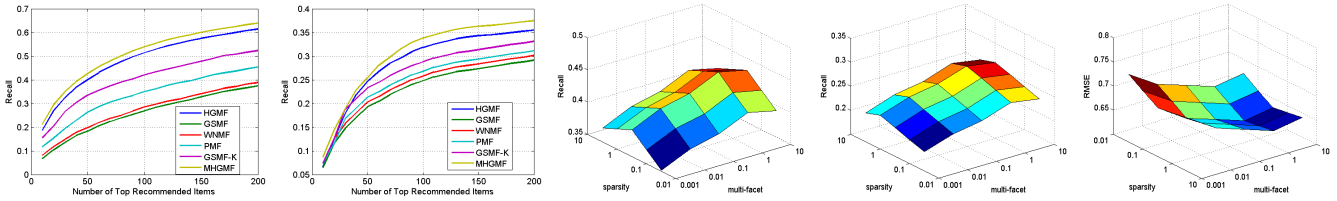
Impact of the Graph Weight and Hidden Group Numbers. Table 4 and Figure 3 show that for the dataset LastFM, higher λ_W suggests good performance. For other two sparser datasets, using relatively lower value will improve the performance. It could be explained that λ_W controls the degree of group membership. In the sparser cases, each user may rely on more neighbours which prefers lower value. In the denser case, it is more likely to introduce more noisy patterns. Thus, the larger regularization will improve the discriminative performance. Table 5 and Figure 4 show that fixing the optimal λ_W, setting larger number of hidden groups will achieve better results. Intuitively, enlarging the number of hidden groups will narrow the range of visible neighbours. When the invisible neighbours are beneficial, the non-overlapping assumption will hurt the performance. In contrast, our method can achieve the same goal of learning discriminative representation but without the restriction. Our results show good and more stable performance in that case, which demonstrates the overlapping assumption is more reasonable for real data.

Impact of Multi-facet Penalty and View Sparsity. Figure 2(c,d,e) shows that in general, when increasing the value of multi-facet penalty λ_1 or view sparsity penalty λ_2, the performance tends to first increase and then decrease. Among λ_1 and λ_2, the performance is relatively sensitive to λ_2. It could be explained that view sparsity controls the discriminative power through automatic latent factor selection. In this view, the proposed method can be also seen an extension of GSMF, due to the similar goals to enhance discriminative power by modeling user interests with different subset of latent factors, but without needing group pre-partition.

Visualization. In addition, we create 2D plots in Figure 5 using MDS [Wang and Boyer, 2013], to illustrate the overlapping decomposition of data on Lastfm dataset.

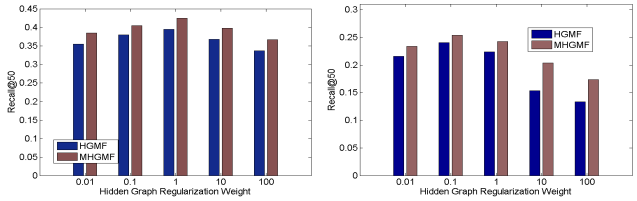
Dataset	Training	WNMF	PMF	GSMF	GSMF-K	HGMF	MHGMF
Epinions	90%	0.6815	0.6732	0.6923	0.6683	0.6447	**0.6209**
	70%	0.6885	0.6789	0.6993	0.6732	0.6482	**0.6313**
	50%	0.7639	0.7510	0.7721	0.7291	0.7140	**0.7004**

Table 3: Performance comparison for RMSE on different sparsity cases. The standard deviations are ≤ 0.01.



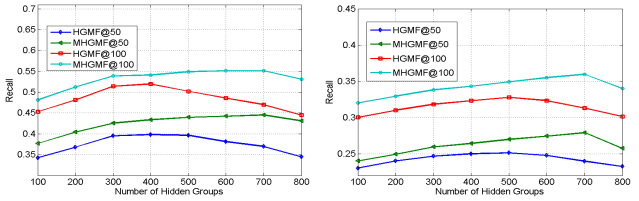
(a) LastFM (b) Delicious (c) LastFM:Recall@50 (d) Delicious:Recall@50 (e) Epinions:RMSE

Figure 2: Subfigure (a),(b): Model comparison for Recall. Subfigure (c),(d),(e): Parameter sensitivity analysis of multi-facet penalty and view sparsity. We show the value of view sparsity regularization parameter λ_2 as $\{0.01, 0.1, 1, 10\}$ and the value of multi-facet regularization parameter λ_1 as $\{0.001, 0.01, 0.1, 1, 10\}$. For Recall, higher is better, and for RMSE, lower is better.



(a) LastFM λ_W (b) Delicious λ_W

Figure 3: Impact of Hidden Graph Weight.

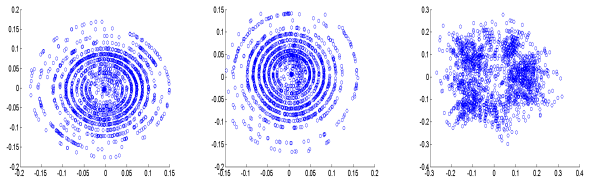


(a) LastFM $\lambda_W = 1$ (b) Delicious $\lambda_W = 0.1$

Figure 4: Impact of the Number of Hidden Groups.

λ_W/RMSE	**0.01**	**0.1**	**1**	**10**	**100**
HGMF	0.6719	0.6447	0.6991	0.7583	0.8321
MHGMF	0.6541	0.6209	0.6474	0.6971	0.7509

Table 4: Impact of Hidden Graph Weight (k=300, Epinions).



(a) View 1 (b) View 2 (c) Combined View

Figure 5: Case study of overlapping decomposition. Our algorithm can capture the original overlapping case, by learning the decomposed non-overlapping multi-facet structures.

5 Related Work

The use of structured information in collaborative filtering is not new [Purushotham and Liu, 2012; Wang and Blei, 2011; Yuan *et al.*, 2014]. Recently, several emerging studies have paid attention to explore implicit or hidden structures for collaborative filtering. [Zhang *et al.*, 2013] proposed a general pipeline framework (LMF), which is independent of specific matrix factorization models. Our model might be further improved by using that framework. [Wang *et al.*, 2014] proposed a hierarchical group matrix factorization method, which needs to obtain the group information in advance, by using side information or pre-clustering. [Wang *et al.*, 2015] proposed an implicit hierarchical matrix factorization approach. It is a direct deeply factorizing approach with fixed tree structures. In contrast, our method focuses on modeling hidden group structures in a graph based regularization framework. [Yuan *et al.*, 2014; Zhang and Wang, 2015] are the most similar works to this paper, which explored modeling hidden group structures by automatic latent factor selection and by Laplacian constraint respectively. However, they either assume the items have already been categorized into multiple semantic groups, or the underlying structures are non-overlapping. In our method, we relax the non-overlapping assumption without requiring pre-grouping in advance, by implicitly seeking the discriminative group-specific latent factors in an adaptive multi-view setting.

6 Conclusion

In this paper, we present a new viewpoint of hidden graph regularization for collaborative filtering, to explicitly model complex hidden structures while learning matrix factorization. We developed a novel structured prior as constraint on the underlying low-rank structures, which generalizes the recently proposed Laplacian constraint via overlapping decomposition, to automatically capture multi-view hidden struc-

#k/RMSE	**100**	**300**	**500**	**700**	**800**
HGMF	0.6796	0.6447	0.6531	0.6603	0.6684
MHGMF	0.6452	0.6209	0.6142	0.6203	0.6291

Table 5: Impact of Hidden Group Numbers (Epinions).

tures, by solving a constraint optimization problem jointly with learning user and item representations. Experiments on real-world datasets exhibit the promising performance, compared with the baseline state-of-the-art methods. Future work could consider to incorporate explicit social graph into the proposed framework, to further improve the performance.

7 Acknowledgements

Our work is supported by National Natural Science Foundation of China (No.61370117 & No.61433015) and Major National Social Science Fund of China (No.12&ZD227). The corresponding author of this paper is Houfeng Wang.

References

[Bach and Jordan, 2002] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[Ding *et al.*, 2006] Chris H. Q. Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R_1-pca: rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *Proceedings of ICML, 2006*, pages 281–288, 2006.

[Feng *et al.*, 2014] Jiashi Feng, Zhouchen Lin, Huan Xu, and Shuicheng Yan. Robust subspace segmentation with block-diagonal prior. In *Proceedings of CVPR*, pages 3818–3825, 2014.

[Gao *et al.*, 2013] Jing Gao, Jiawei Han, Jialu Liu, and Chi Wang. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 13th SIAM International Conference on Data Mining*, pages 252–260, 2013.

[Golub and van Loan, 1996] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of ALT, 2005*, pages 63–77, 2005.

[Gu *et al.*, 2011] Quanquan Gu, Zhenhui Li, and Jiawei Han. Joint feature selection and subspace learning. In *Proceedings of IJCAI, 2011*, pages 1294–1299, 2011.

[Hu *et al.*, 2008] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of ICDM, December 15-19, 2008, Pisa, Italy*, pages 263–272, 2008.

[Kumar and III, 2011] Abhishek Kumar and Hal Daumé III. A co-training approach for multi-view spectral clustering. In *Proceedings of ICML*, pages 393–400, 2011.

[Li *et al.*, 2011] Mu Li, Xiao-Chen Lian, James T. Kwok, and Bao-Liang Lu. Time and space efficient spectral clustering via column sampling. In *CVPR*, pages 2297–2304, 2011.

[Lin *et al.*, 2011] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Proceedings of NIPS*, pages 612–620, 2011.

[Masaeli *et al.*, 2010] Mahdokht Masaeli, Glenn Fung, and Jennifer G. Dy. From transformation-based dimensionality reduction to feature selection. In *Proceedings of ICML*, pages 751–758, 2010.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of NIPS*, pages 849–856, 2001.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint l21-norms minimization. In *Proceedings of NIPS*, pages 1813–1821, 2010.

[Niu *et al.*, 2010] Donglin Niu, Jennifer G. Dy, and Michael I. Jordan. Multiple non-redundant spectral clustering views. In *Proceedings of ICML*, pages 831–838, 2010.

[Purushotham and Liu, 2012] Sanjay Purushotham and Yan Liu. Collaborative topic regression with social matrix factorization for recommendation systems. In *Proceedings of ICML*, 2012.

[Salakhutdinov and Mnih, 2007] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Proceedings of NIPS*, pages 1257–1264, 2007.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Trendafilov, 2010] Nickolay T. Trendafilov. P.-A. absil, r. mahony, and r. sepulchre. optimization algorithms on matrix manifolds. *Foundations of Computational Mathematics*, 10(2):241–244, 2010.

[von Luxburg, 2007] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[Wang and Blei, 2011] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of KDD*, pages 448–456, 2011.

[Wang and Boyer, 2013] Quan Wang and Kim L. Boyer. Feature learning by multidimensional scaling and its applications in object recognition. *CoRR*, abs/1306.3294, 2013.

[Wang *et al.*, 2014] Xin Wang, Weike Pan, and Congfu Xu. HGMF: hierarchical group matrix factorization for collaborative recommendation. In *Proceedings of CIKM*, pages 769–778, 2014.

[Wang *et al.*, 2015] Suhang Wang, Jiliang Tang, Yilin Wang, and Huan Liu. Exploring implicit hierarchical structures for recommender systems. In *Proceedings of IJCAI*, pages 1813–1819, 2015.

[Yuan *et al.*, 2014] Ting Yuan, Jian Cheng, Xi Zhang, Shuang Qiu, and Hanqing Lu. Recommendation by mining multiple user behaviors with group sparsity. In *Proceedings of AAAI*, pages 222–228, 2014.

[Zhang and Wang, 2015] Qing Zhang and Houfeng Wang. Improving collaborative filtering via hidden structured constraint. In *Proceedings of CIKM*, pages 1935–1938, 2015.

[Zhang *et al.*, 2006] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of SDM*, pages 549–553, 2006.

[Zhang *et al.*, 2013] Yongfeng Zhang, Min Zhang, Yiqun Liu, Shaoping Ma, and Shi Feng. Localized matrix factorization for recommendation based on matrix block diagonal forms. In *Proceedings of WWW*, pages 1511–1520, 2013.