Towards Robust Co-Clustering

Liang Du^{1,2} and **Yi-Dong Shen**^{1}
^{1}State Key Laboratory of Computer Science,
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
^{2}University of Chinese Academy of Sciences, Beijing 100049, China
{duliang,ydshen}@ios.ac.cn

Abstract

Nonnegative Matrix Tri-factorization (NMTF) and its graph regularized extensions have been widely used for co-clustering task to group data points and features simultaneously. However existing methods are sensitive to noises and outliers which is because of the squared loss function is used to measure the quality of data reconstruction and graph regularization. In this paper, we extend GNMTF by introducing a sparse outlier matrix into the data reconstruction function and applying the ℓ_1 norm to measure graph dual regularization errors, which leads to a novel Robust Co-Clustering (RCC) method. Accordingly, RCC is expected to obtain a more faithful approximation to the data recovered from sparse outliers, and achieve robust regularization by reducing the regularization errors of unreliable graphs via ℓ_1 norm. To solve the optimization problem of RCC, an alternating iterative algorithm is provided and its convergence is also proved. We also show the connection between the sparse outlier matrix in data reconstruction function and the robust Huber M-estimator. Experimental results on real-world data sets show that our RCC consistently outperforms the other algorithms in terms of clustering performance, which validates the effectiveness and robustness of the proposed approach.

1 Introduction

Clustering is a fundamental topic in unsupervised machine learning. To group similar data points into clusters, numerous clustering algorithms have been developed, e.g., Kmeans, spectral clustering and NMF [Seung and Lee, 2001]. However, these methods only consider data side clustering and omit the cluster structure on feature side. Typically, the clustering of samples and features are closely related in real world applications.

Motivated by the duality and interdependence between data and feature clusters, several co-clustering approaches have been proposed to partition both dimensions simultaneously by modeling the relationship between samples and features, such as graph-based method [Dhillon, 2001], information-theoretic based method [Dhillon *et al.*, 2003] and matrix factorization based method [Ding *et al.*, 2006]. More recently, it has been shown that the intersample and interfeature relationships also play important roles in co-clustering task [Gu and Zhou, 2009; Zhang *et al.*, 2012b]. [Gu and Zhou, 2009; Shang *et al.*, 2012] proposed dual regularized (semi)-GNMTF models, which impose graph regularization on both data and feature cluster assignment matrices, to explore the geometric structure of data manifold and feature manifold. [Gu *et al.*, 2011] further imposed a normalized cut like constraint to make GNMTF model well-defined. [Wang *et al.*, 2011] proposed a fast GNMTF method by constraining factors to be cluster indicator matrices. Besides, GNMTF method has also been developed to address other aspect of co-clustering [Long *et al.*, 2012].

Although GNMTF based co-clustering methods have attracted increasing attention in recent years, they still surffer from the following problems in real-world applications. First, these methods adopt the squared loss, which is well-known to be unstable with respect to outliers and noises, to measure the reconstruction error incurred by modeling the relationship between sample and features. Accordingly, a few noisy entries with large errors may easily dominate the factorization. Second, the intersample and interfeature relationships, such as the regularization on data and feature graphs in GNMTF, are also essential for co-clustering task. When the data set is corrupted by outliers and noises, the constructed graphs may also be contaminated [Zhang *et al.*, 2012a] and fail to capture the desired geometrical information, for instance, two samples may be mis-linked and incurred inappropriate weight in a nearest neighbor graph. As a result, the graph regularization may mislead the factorization and deteriorate the performance.

In this paper, a novel Robust Co-Clustering (RCC) method is presented by extending GNMTF towards algorithmic robustness to both data outliers and unreliable graphs. Specifically, to model the sample-feature relationship faithfully, we introduce a sparse error matrix to explicitly capture the grossly corrupted entries in the data reconstruction function. With this error matrix, a much cleaned data could be recovered and a robust factorization result could be expected. Besides, to robustly model the intersample and iterfeature relationships, we apply the ℓ_1-norm error function, instead of the traditional ℓ_2 function, to alleviate the influence of large and unexpected regularization incurred by unreliable graphs. In this way, the error matrix in data reconstruction function alleviates the influence of noisy data, and the ℓ_1-norm reduces the impact of unreliable graph regularization errors and leads to sparse and faithful regularization. Hence, RCC is robust to noisy data and unreliable graphs. To solve the optimization problem of our approach, we derive an alternating iterative algorithm and prove its convergence. We also show the connection between the error matrix in data reconstruction function and the Huber M-estimator in robust statistics [Hampel *et al.*, 2011]. Thus, the robustness of our data reconstruction function can be interpreted as reducing the influence of large reconstruction errors which are often caused by outlier entries. Experimental results on real data sets show that RCC consistently outperforms the other algorithms in terms of clustering results, and possess robustness to noisy data and unreliable graphs.

2 Preliminaries

Throughout the paper, we use small letters (e.g., a) to denote scalars, capital letters (e.g., A) to denote matrices. We denote $A_{i\cdot}, A_{\cdot j}$, and A_{ij} as the i-th row, the j-th column and the entry at the i-th row and j-th column of a matrix A. We use $\mathrm{tr}(A)$ to denote the trace of a square matrix A, $\odot$ to denote the element-wise product and $\oslash$ for element-wise division. We also define $\mathbf{1}_d \in \mathbb{R}^d$ as a column vector with all its elements equals to 1.

Given a nonnegative data matrix $X \in \mathbb{R}^{d\times n}$ where rows correspond to features and columns to samples, the goal of co-clustering is to group the features $\{X_{1\cdot}, \ldots, X_{d\cdot}\}$ into k_1 clusters $\{\mathcal{C}_i^{'}\}_{i=1}^{k_1}$, while group the data points $\{X_{\cdot 1}, \ldots, X_{\cdot n}\}$ into k_2 clusters $\{\mathcal{C}_i\}_{i=1}^{k_2}$.

2.1 Graph Regularized NMTF

To co-cluster data points and features simultaneously, [Gu and Zhou, 2009; Shang *et al.*, 2012] proposed graph regularized (semi-)NMTF models by imposing graph regularization on data and feature graphs under the manifold assumption, which can be formulated as minimizing the following objective function

$$\begin{aligned}\mathcal{J} = &||X - FHG^T||_F^2 \\ &+ \lambda_F \mathrm{tr}(F^T L_F F) + \lambda_G \mathrm{tr}(G^T L_G G) \\ &\text{s.t.} \quad F \geq 0, H \geq 0, G \geq 0,\end{aligned} \tag{1}$$

where F and G are feature and data point cluster assignments matrices, H reflects the association between feature clusters and data clusters, $L_F = D_F - W_F$ is the graph Laplacian on the feature graph, D_F and W_F are the degree matrix and the adjacency matrix of the feature graph. The definition of L_G for data graph is the same as above, λ_F and λ_G are regularization parameters.

3 Robust Co-Clustering

3.1 Formulation

Motivated by the recent development in robust principle component analysis [Cands *et al.*, 2011], we propose a novel nonnegative matrix tri-factorization model to approximate the data matrix, where we assume some entries of the data matrix may be arbitrarily corrupted, but the corruption is sparse. We introduce an error matrix $S \in \mathbb{R}^{d\times n}$ to explicit capture the sparse corruption. Thus, the goal of robust tri-factorization is to approximate the nonnegative data matrix X as

$$X \approx FHG^T + S,$$

where $F \in \mathbb{R}^{d\times k_1}, H \in \mathbb{R}^{k_1\times k_2}$ and $G \in \mathbb{R}^{n\times k_2}$ are constrained to be nonnegative, and S is required to be sparse. Due to the presence of S, a much clean data matrix can be recovered from sparse corruption, and more faithful factorization result could be expected.

Besides, the intersample and interfeature relationships, such as the dual graph regularizers in GNMTF, are also essential for clustering task [Zhang *et al.*, 2012b]. However, when the data set is corrupted by outliers and noises, the constructed graph may be contaminated and not reliable to capture the desired geometrical information. For instance, two data points are mis-linked in a KNN graph due to outliers. As a result, the graph regularization may mislead the factorization and deteriorate the performance. Considering two far away samples or features are forced to be close according to unreliable graph regularization, we further propose to minimize the ℓ_1-norm regularization error which can suppress the large and unexpected regularization errors, inspired by recent developments in graph embedding [Zhang *et al.*, 2012a]. As a result, the affect of large regularization errors caused by unreliable graphs can be reduced. By minimizing the ℓ_1-norm of graph regularization error, we can also re-estimate more faithful graph structure such as the weights between samples or features as shown later.

By combining robust factorization and robust dual graph regularization into a unified framework, our Robust Co-Clustering (RCC) method can be formulated as follows

$$\begin{aligned}\mathcal{J}_1 = &||X - FHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 \\ &+ \lambda_S ||S||_1 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2 \\ &\text{s.t.} \quad F, H, G \geq 0, F\mathbf{1}_{k1} = \mathbf{1}_d, G\mathbf{1}_{k2} = \mathbf{1}_n,\end{aligned} \tag{2}$$

where λ_S, λ_F and λ_G are regularization parameters. With the ℓ_1 normalization constraints on each row of F and G, and the ℓ_1 norm of regularization errors in the last two terms, the optimization problem in Eq. (2) is well defined and no longer suffers from either scale transfer problem or trivial solution in GNMTF [Gu *et al.*, 2011].

3.2 Algorithm

The problem in Eq. (2) is not convex in all variables together, but convex in them separately. In the following, we introduce an iterative algorithm based on block coordinate descent to minimize the objective function in Eq. (2). We separately update the value of S, F, H, and G while holding the other variables as constant. Thus, a local minima can be expected by solving a sequence of convex optimization problems. We also prove the convergence of the algorithm.

Update S

Optimizing Eq. (2) w.r.t. S is equivalent to minimizing

$$\mathcal{J}_2 = ||X - FHG^T - S||_F^2 + \lambda_S ||S||_1. \tag{3}$$

It is easy to show that the above problem is element-wise decoupled, and the unique solution of each subproblem admits a closed form called the soft-thresholding operator [Bach *et al.*, 2012], i.e.,

$$S_{ij} = \begin{cases} 0 & \text{if } |E_{ij}| \le \lambda_S/2, \\ E_{ij} - \frac{\lambda_S}{2}\mathrm{sign}(E_{ij}) & \text{otherwise,} \end{cases} \tag{4}$$

where $E_{ij} = (X - FHG^T)_{ij}$. By substituting Eq. (4) into Eq. (3), we get

$$\min \mathcal{J}_2 = \begin{cases} E_{ij}^2 & \text{if } |E_{ij}| \le \lambda_S/2, \\ \lambda_S|E_{ij}| - (\frac{\lambda_S}{2})^2 & \text{otherwise.} \end{cases} \tag{5}$$

It is interesting to note that the right part of Eq. (5) is often denoted as Huber M-estimator in robust statistics [Hampel *et al.*, 2011]. Based on the duality between Eq. (3) and the Huber loss in Eq. (5), it can be seen that: if $\lambda_S > 2\max_{ij}|E_{ij}|$, Eq. (3) degenerates to ordinary NMTF; if $\lambda_S \to 0$, Eq. (3) behaves like ℓ_1-norm factorization; Eq. (3) behaves like ℓ_2-norm on small errors ($E_{ij} \le \lambda_S/2$) and like ℓ_1-norm on large errors by a trade off parameter λ_S. Different from traditional methods [Shang *et al.*, 2012; Gu *et al.*, 2011] which use the squared loss on all reconstruction errors, our method automatically adapts the ℓ_1-norm on large errors, which are often caused by outliers. Thus, the influence of outliers with large errors can be alleviated. Besides, the reformulation in Eq. (5) can also be used to guide the selection of the regularization parameter λ_S.

Update F

Optimizing Eq. (2) w.r.t. F is equivalent to minimizing

$$\begin{aligned} \mathcal{J}_3 = ||X - FHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 \\ \text{s.t.} \quad F \ge 0, F\mathbf{1}_{k_1} = \mathbf{1}_d. \end{aligned} \tag{6}$$

To deal with the nonsmoothness of the ℓ_1-norm on graph regularization, we first apply the quadratic variational form for the ℓ_1-norm [Bach *et al.*, 2012], so that we can rewrite Eq. (6) in a compact form:

$$\begin{aligned} \mathcal{J}_4 = -2\mathrm{tr}(F^TP^+) + 2\mathrm{tr}(F^TP^-) + \mathrm{tr}(FQF^T) \\ + \lambda_F \mathrm{tr}(F^T\widetilde{D}^FF) - \lambda_F\mathrm{tr}(F^T\widetilde{W}^FF) \\ \text{s.t.} \quad F \ge 0, F\mathbf{1}_{k_1} = \mathbf{1}_d, \end{aligned} \tag{7}$$

where $\widetilde{W}_{ii'}^F = \frac{W_{ii'}^F}{2||F_{i\cdot} - F_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^F = \sum_{i'}\widetilde{W}_{ii'}^F$, $P = (X - S)GH^T$, $Q = HG^TGH^T$, and we introduce $P_{ij}^+ = (|P_{ij}| + P_{ij})/2$, $P_{ij}^- = (|P_{ij}| - P_{ij})/2$.

Note that although the ℓ_1-norm constraints on latent factors (F and G) have been explored in NMTF based methods in previous works [Zhuang *et al.*, 2011; Long *et al.*, 2012], the constrained optimization problem is usually solved by a two-step strategy, which first minimizes an unconstrained problem without considering the constraints and then uses an additional ℓ_1 normalization operator on each row of F and G in each iteration. However, this is not a principled way to solve the problem, and the price of normalization is potential performance decrease [Sun and Hamme, 2012].

Inspired from [Sun and Hamme, 2012], we first project F onto the ℓ_1 hyperplane by introducing a nonnegative factor matrix $\widetilde{F} \in \mathbb{R}^{d \times k_1}$, the row normalized F can be obtained by $F_{ik} = \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$. Thus, we can optimize Eq. (7) by minimizing the following objective function with respect to $\widetilde{F}$

$$\begin{aligned} \mathcal{J}_5(\widetilde{F}) = -2\sum_{ik} P_{ik}^+ \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + 2\sum_{ik} P_{ik}^- \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + \\ \sum_{ikk'} Q_{kk'} \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} \frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}} + \lambda_F \sum_{ii'k} (\widetilde{L}_F)_{ii'}^+ \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} \frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}} \\ - \lambda_F \sum_{ii'k} (\widetilde{L}_F)_{ii'}^- \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} \frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}}, \end{aligned} \tag{8}$$

where the solution of F derived from $\widetilde{F}$ always satisfies the ℓ_1 normalization constraints.

The auxiliary function [Gu and Zhou, 2009] of the above objective function can be constructed as follows

$$\begin{aligned} Z(\widetilde{F}, \widetilde{F}^t) = -2\sum_{ik} P_{ik}^+ \frac{\widetilde{F}_{ik}^t}{\sum_s \widetilde{F}_{is}^t} (1 + \log \frac{\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is}}{\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t}) \\ + 2\sum_{ik} P_{ik}^- \frac{(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2 + (\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t)^2}{2\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t} \\ + \sum_{ik} \frac{[(\widetilde{F}^t \oslash (\widetilde{F}\mathbf{1}_{k_1}\mathbf{1}_{k_1}^T))Q]_{ik}(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2}{\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t} \\ + \lambda_F \sum_{ik} \frac{[\widetilde{D}^F(\widetilde{F}^t \oslash (\widetilde{F}\mathbf{1}_{k_1}\mathbf{1}_{k_1}^T))]_{ik}(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2}{\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t} \\ - \lambda_F \sum_{ii'k} \widetilde{W}_{ii'}^F \left\{ (\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t)(\widetilde{F}_{i'k}^t/\sum_{k'} \widetilde{F}_{i'k'}^t) \right. \\ \left. (1 + \log \frac{(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})(\widetilde{F}_{i'k}/\sum_{k'} \widetilde{F}_{i'k'})}{(\widetilde{F}_{ik}^t/\sum_s \widetilde{F}_{is}^t)(\widetilde{F}_{i'k}^t/\sum_{k'} \widetilde{F}_{i'k'}^t)}) \right\}. \end{aligned} \tag{9}$$

Setting the partial derivative of $\widetilde{F}_{ik}$ to zero, we have

$$\frac{\partial Z(\widetilde{F}, \widetilde{F}^t)}{\partial \widetilde{F}_{ik}} = \sum_{k'} \frac{\partial Z(\widetilde{F}, \widetilde{F}^t)}{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})} \frac{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}{\partial \widetilde{F}_{ik}} = 0. \tag{10}$$

Since $F_{ik} = \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$, then we have

$$\begin{aligned} \frac{\partial Z(\widetilde{F}, \widetilde{F}^t)}{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})} = -2\left\{P_{ik'}^+ + \lambda_F[\widetilde{W}^FF^t]_{ik'}\right\} \frac{F_{ik'}^t}{F_{ik'}} \\ + 2\left\{P_{ik'}^- + [FQ]_{ik'} + \lambda_F[\widetilde{D}^FF^t]_{ik'}\right\} \frac{F_{ik'}}{F_{ik'}^t}, \end{aligned} \tag{11}$$

$$\frac{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}{\partial \widetilde{F}_{ik}} = \frac{\delta_{ik'} - F_{ik'}}{\sum_s \widetilde{F}_{is}}, \delta_{ik'} = \begin{cases} 1 & \text{if } k = k' \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Let $A = (P^- + F^t Q + \lambda_F \widetilde{D}^F F^t) \oslash F^t$, $C = (P^+ + \lambda_F \widetilde{W}^F F^t) \odot F^t$, then Eq. (10) is equivalent to

$$A_{ik} F_{ik}^2 + \sum_s [C_{is} - A_{is} F_{is}^2] F_{ik} - C_{ik} = 0. \tag{13}$$

Note that only F is involved in the Eq. (13). Summing Eq. (13) over k, we have

$$\sum_k A_{ik} F_{ik}^2 + \sum_s [C_{is} - A_{is} F_{is}^2] \sum_k F_{ik} - \sum_k C_{ik}$$
$$= \sum_s [C_{is} - A_{is} F_{is}^2] (\sum_k F_{ik} - 1) = 0. \tag{14}$$

It is obvious that the solution of Eq (14) always satisfies the ℓ_1 normalization constraint for almost all λ_F.

In this paper, we use Algorithm 1 developed in [Sun and Hamme, 2012] to compute the fixed point of Eq. (13), where only element-wise operations are involved and a few iterations are needed to converge (i.e, $nIter$ = 10).

Algorithm 1 Algorithm to compute the fixed point of Eq.(13)

```
Input: $A$, $C$, $F^t$, $nIter$ and $k$.
Output: $F$
  while $iter \leq nIter$ do
    $B = (C - A \otimes F^2) \mathbf{1}_k \mathbf{1}_k^T$;
    $F = (\sqrt{B^2 + 4AC} - B) \oslash (2A)$
    $\mathcal{I} = \{j | B_{j1} < 0\}$
    if $\mathcal{I}$ is not empty then
      $F_{\mathcal{I}\cdot} = \mathrm{diag}(F_{\mathcal{I}\cdot} \mathbf{1}_k)^{-1} F_{\mathcal{I}\cdot}$
    end if
  end while
```

Update H

Optimizing Eq. (2) w.r.t. H is equivalent to minimizing

$$\mathcal{J}_5 = ||X - FHG^T - S||_F^2, \quad \text{s.t.} \quad H \geq 0. \tag{15}$$

Introducing the Lagrangian multiplier $\Lambda_H \in \mathbb{R}^{k_1 \times k_2}$ and setting the partial derivative to zero, we obtain

$$\Lambda_H = -2F^T (X - S) G + 2F^T FHG^T G. \tag{16}$$

Using the Karush-Kuhn-Tucker condition $(\Lambda_H)_{ij} H_{ij}^2 = 0$, we get the following updating rule

$$H_{ij} = H_{ij} \sqrt{\frac{[F^T (X - S) G]_{ij}^+}{[F^T FHG^T G]_{ij} + [F^T (X - S) G]_{ij}^-}}. \tag{17}$$

Update G

Optimizing Eq. (2) w.r.t. G is equivalent to minimizing

$$\mathcal{J}_6 = ||X - FHG^T - S||_F^2 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2$$
$$\text{s.t.} \quad G \geq 0, G\mathbf{1}_{k_2} = \mathbf{1}_n. \tag{18}$$

Considering the duality between F and G, the minimization of Eq. (18) w.r.t. G can also be solved by the algorithm developed for optimizing F in Eq. (6).

In summary, we present the alternating iterative algorithm for optimizing Eq. (2) in Algorithm 2.

Algorithm 2 Algorithm to solve the problem in Eq. (2)

```
Input: The data matrix $X$, the initial values of $F, H, G$,
  data and feature graphs $W_F, W_G$, regularization parame-
  ters $\lambda_S, \lambda_F, \lambda_G$.
Output: Partition matrices $F$ and $G$.
  repeat
    Update $S$ according to Eq. (4);
    Let $\widetilde{W}_{ii'}^F = \frac{W_{ii'}^F}{2||F_{i\cdot} - F_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^F = \sum_{i'} \widetilde{W}_{ii'}^F$, $P = (X - S) G H^T$, $Q = H G^T G H^T$;
    Set $A = (P^- + FQ + \lambda_F \widetilde{D}^F F) \oslash F$;
    Set $C = (P^+ + \lambda_F \widetilde{W}^F F) \odot F$;
    Update $F$ according to Algorithm 1;
    Update $H$ according to Eq. (17);
    Let $\widetilde{W}_{ii'}^G = \frac{W_{ii'}^G}{2||G_{i\cdot} - G_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^G = \sum_{i'} \widetilde{W}_{ii'}^G$, $P = (X - S)^T F H$, $Q = H^T F^T F H$;
    Set $A = (P^- + GQ + \lambda_G \widetilde{D}^G G) \oslash G$;
    Set $C = (P^+ + \lambda_G \widetilde{W}^G G) \odot G$;
    Update $G$ according to Algorithm 1;
  until Converges
```

3.3 Convergence Analysis

In this subsection, we prove the convergence of the proposed algorithm by using the properties of auxiliary function developed in [Seung and Lee, 2001] and the following Lemma provided in [Nie *et al.*, 2010].

Lemma 1. *For any non-zeros vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, the following results follows [Nie* et al.*, 2010]*

$$||\boldsymbol{a}||_2 - \frac{||\boldsymbol{a}||_2^2}{2||\boldsymbol{b}||_2} \leq ||\boldsymbol{b}||_2 - \frac{||\boldsymbol{b}||_2^2}{2||\boldsymbol{b}||_2}. \tag{19}$$

Theorem 1. *The Algorithm 2 monotonically decreases the objective function of RCC in Eq.* (2) *in each iteration and converges to an optimal solution.*

Proof. Since the objective function of RCC in Eq. (2) is bounded below, we only need to prove the objective function monotonically decreases under the updates of $S^{t+1}, F^{t+1}, H^{t+1}$, and G^{t+1} in each iteration. The solution of S^{t+1} in Eq. (4) is the global minima of the subproblem in Eq. (3). Based on the properties of the auxiliary function [Seung and Lee, 2001], we have $\mathcal{J}_4(F^t) \geq Z(F^{t+1}, F^t) \geq Z(F^{t+1}, F^{t+1}) = \mathcal{J}_4(F^{t+1})$. Thus the objective function in Eq. (7) will monotonically decrease under the update of F^{t+1}, that is $\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_4(F^t)$, i.e.,

$$||X - F^{t+1} HG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F \frac{||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2}$$
$$\leq ||X - F^t HG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F \frac{||F_{i\cdot}^t - F_{i'\cdot}^t||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2}.$$

So, we only need to prove the update of F^{t+1} satisfies $\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_3(F^t)$. And we have the following inequality based on $\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_4(F^t)$

$$\begin{aligned}
&||X - F^{t+1}HG^T - S||_F^2 + \lambda_F \sum_{ii'} W^F_{ii'} ||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2 \\
&+ \lambda_F \sum_{ii'} W_{ii'} \left(\frac{||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2^2}{2||F^t_{i\cdot} - F^t_{i'\cdot}||_2} - ||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2 \right) \\
\leq\, &||X - F^tHG^T - S||_F^2 + \lambda_F \sum_{ii'} W^F_{ii'} ||F^t_{i\cdot} - F^t_{i'\cdot}||_2 \\
&+ \lambda_F \sum_{ii'} W_{ii'} \left(\frac{||F^t_{i\cdot} - F^t_{i'\cdot}||_2^2}{2||F^t_{i\cdot} - F^t_{i'\cdot}||_2} - ||F^t_{i\cdot} - F^t_{i'\cdot}||_2 \right) \quad (20)
\end{aligned}$$

Recalling the result in Lemma 1, we know that

$$\begin{aligned}
&\frac{||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2^2}{2||F^t_{i\cdot} - F^t_{i'\cdot}||_2} - ||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2 \\
&\qquad \geq \frac{||F^t_{i\cdot} - F^t_{i'\cdot}||_2^2}{2||F^t_{i\cdot} - F^t_{i'\cdot}||_2} - ||F^t_{i\cdot} - F^t_{i'\cdot}||_2. \quad (21)
\end{aligned}$$

Combining Eq. (20) with Eq. (21), we have the following results

$$\begin{aligned}
&||X - F^{t+1}HG^T - S||_F^2 + \lambda_F \sum_{ii'} W^F_{ii'} ||F^{t+1}_{i\cdot} - F^{t+1}_{i'\cdot}||_2 \\
\leq\, &||X - F^tHG^T - S||_F^2 + \lambda_F \sum_{ii'} W^F_{ii'} ||F^t_{i\cdot} - F^t_{i'\cdot}||_2
\end{aligned}$$

This inequality indicates that the objective function $\mathcal{J}_3$ in Eq. (6) will monotonically decrease in each iteration. Similarly, we can also prove the objective function $\mathcal{J}_6$ in Eq. (18) monotonically decreases under the update of G^{t+1}. The non-increasing of the update of H^{t+1} in Eq. (17) can be proved according to the properties of the auxiliary function [Seung and Lee, 2001]. □

4 Experiments

In this section, we evaluate the effectiveness and robustness of our proposed RCC method for clustering. We compared RCC with the following clustering and co-clustering methods: Kmeans, NMF [Seung and Lee, 2001], ONMTF [Ding *et al.*, 2006], GNMF [Cai *et al.*, 2011], DRCC [Gu and Zhou, 2009], DNMTF [Shang *et al.*, 2012], LDCC [Zhang *et al.*, 2012b].

4.1 Parameters Setting

Since each clustering algorithm has one or more parameters to be tuned, in order to compare these algorithms fairly, we run these algorithms under different parameter settings and select the best average result for comparison. We set the number of clusters equal to the true number of classes for all the data sets and clustering algorithms. For graph regularized matrix factorization methods, the neighborhood size p to construct the graph is searched over the grid $\{1, 2, \ldots, 10\}$ and the binary weighting scheme is used for its simplicity. The regularization parameters are searched over the grid $\{0.1, 1, 10, 100, 500, 1000\}$ according to [Gu and Zhou, 2009]. For dual graph regularized methods, the regularization parameters on data and feature graphs are set to the same value. For tri-factorization methods, the number of feature clusters is set to be the same as the number of data clusters. For LDCC, the local regularization parameter is set to be 1 and the embedding dimension is searched over $\{5, 10, \ldots, 50\}$ according to [Zhang *et al.*, 2012b]. Based on the duality between Eq. (3) and the Huber loss in Eq. (5), the regularization parameter λ_S is empirically set to $\lambda_S = 2\text{median}_{ij}(E_{ij})$. Since the reconstruction error E_{ij} is changed in each iteration and we also adjust λ_S accordingly.

4.2 Data Sets

In our experiment, we have collected five public datasets to show the effectiveness of the proposed method. These data sets include one face image dataset, i.e., JAFFE, two hand written digit image datasets from UCI, i.e, MFEA and OPTDIGIT [Shang *et al.*, 2012], two gene expression datasets, i.e., LUNG and GLIOMA, we use the processed data from [Nie *et al.*, 2010]. Table 1 summarizes the characteristics of the data sets used in this experiment.

Table 1: Description of the data sets

Data sets	# samples	# features	# classes
JAFFE	213	676	10
MFEA	2000	240	10
OPTDIGIT	3823	64	10
LUNG	203	3312	5
GLIOMA	50	4434	4

4.3 Evaluation Metrics

The result is evaluated by comparing the cluster label of each sample with the label provided by the data set. We use three metrics to evaluate the clustering performance. One metric is accuracy (ACC), which is used to measure the percentage of correct labels. The second metric is the normalized mutual information (NMI). In clustering applications, mutual information is used to measure how similar two sets of clusters are. The cluster purity is also used to measure the extent to which each cluster contained data points from primarily one class. The definition of these metrics can be found in [Ding *et al.*, 2006; Gu and Zhou, 2009].

4.4 Clustering Results

Under each parameter setting of each method mentioned above, we independently repeat the experiments for 10 times and report the best average result in Table 2. We can see that RCC consistently outperforms the other algorithms in terms of ACC, NMI and Purity on all datasets. This may imply that real world datasets may contain noises and outliers to some extent due to the conditions under which the data sets are captured. By explicitly considering these unexpected corruptions, our method can improve the clustering results in general cases.

5 Conclusion

In this paper, we propose a novel Robust Co-Clustering (RCC) method by extending GNMTF towards robustness to

Table 2: Clustering results measured by Accuracy/NMI/Purity of the compared methods.

Data	Metrics	Kmeans	NMF	ONMTF	GNMF	DRCC	DNMTF	LDCC	RCC
JAFFE	ACC	0.7753	0.8324	0.8178	0.9352	0.9347	0.9535	0.8465	**0.9681**
	NMI	0.8280	0.8314	0.8313	0.9402	0.9363	0.9592	0.8904	**0.9654**
	Purity	0.8038	0.8350	0.8245	0.9413	0.9408	0.9587	0.8634	**0.9681**
MFEA	ACC	0.6908	0.6713	0.6972	0.8934	0.8851	0.9233	0.8713	**0.9237**
	NMI	0.6981	0.6223	0.6266	0.8499	0.8746	0.8746	0.8670	**0.8839**
	Purity	0.7276	0.6928	0.7137	0.8989	0.9178	0.9240	0.8826	**0.9378**
OPTDIGIT	ACC	0.7459	0.7431	0.7200	0.8569	0.8643	0.8674	0.8555	**0.8772**
	NMI	0.7208	0.6779	0.6631	0.8484	0.8642	0.8659	0.8641	**0.8758**
	Purity	0.7659	0.7526	0.7369	0.8577	0.8729	0.8704	0.8741	**0.8775**
LUNG	ACC	0.7215	0.6318	0.7670	0.8754	0.8828	0.8985	0.8158	**0.9281**
	NMI	0.5357	0.4964	0.5507	0.6713	0.6728	06926	0.5754	**0.7296**
	Purity	0.8692	0.8845	0.8988	0.9123	0.9015	0.9232	0.8970	**0.9320**
GLIOMA	ACC	0.5848	0.5870	0.5950	0.6340	0.6560	0.6140	0.6060	**0.6840**
	NMI	0.4621	0.4190	0.4907	0.4875	0.5074	0.4936	0.3636	**0.5350**
	Purity	0.6090	0.4790	0.6270	0.6540	0.6600	0.6440	0.6180	**0.6880**

noisy data and unreliable graphs. Specifically, we introduce a sparse outlier matrix into the data reconstruction function and adopt the ℓ_1 norm to measure the graph regularization errors. As a result, the proposed method can achieve robust factorization by approximating the cleaned data recovered from sparse outliers, and achieve robust regularization by suppressing the large regularization errors of unreliable graphs via ℓ_1 norm. One possible extension is to devise robust GNMTF model under the half-quadratic minimization framework which is helpful in general cases.

6 Acknowledgments

This work is supported in part by NSFC grant 60970045 and China National 973 project 2013CB329305.

References

[Bach *et al.*, 2012] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *FTML*, pages 1–106, 2012.

[Cai *et al.*, 2011] D. Cai, X. He, J. Han, and T.S. Huang. Graph regularized nonnegative matrix factorization for data representation. *PAMI*, 33(8):1548–1560, 2011.

[Cands *et al.*, 2011] E. Cands, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *JACM*, 2011.

[Dhillon *et al.*, 2003] I.S. Dhillon, S. Mallela, and D.S. Modha. Information-theoretic co-clustering. In *SIGKDD*, 2003.

[Dhillon, 2001] I.S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, 2001.

[Ding *et al.*, 2006] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *SIGKDD*, 2006.

[Gu and Zhou, 2009] Q. Gu and J. Zhou. Co-clustering on manifolds. In *SIGKDD*, 2009.

[Gu *et al.*, 2011] Q. Gu, C. Ding, and J. Han. On trivial solution and scale transfer problems in graph regularized nmf. In *IJCAI*, 2011.

[Hampel *et al.*, 2011] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. 2011.

[Long *et al.*, 2012] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang. Transfer learning with graph co-regularization. In *AAAI*, 2012.

[Nie *et al.*, 2010] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *NIPS*, 23:1813–1821, 2010.

[Seung and Lee, 2001] D. Seung and L. Lee. Algorithms for non-negative matrix factorization. *NIPS*, 2001.

[Shang *et al.*, 2012] F. Shang, L.C. Jiao, and F. Wang. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition*, 2012.

[Sun and Hamme, 2012] M. Sun and H. Hamme. Large scale graph regularized nonnegative matrix factorization with ℓ_1 normalization based on kullbackcleibler divergence. In *IEEE TSP*, 2012.

[Wang *et al.*, 2011] H. Wang, F. Nie, H. Huang, and F. Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *IJCAI*, 2011.

[Zhang *et al.*, 2012a] H. Zhang, Z.J. Zha, S. Yan, M. Wang, and T.S. Chua. Robust non-negative graph embedding: Towards noisy data, unreliable graphs, and noisy labels. In *CVPR*, pages 2464–2471, 2012.

[Zhang *et al.*, 2012b] L. Zhang, C. Chen, J. Bu, Z. Chen, D. Cai, and J. Han. Locally discriminative coclustering. *TKDE*, 24(6):1025–1035, 2012.

[Zhuang *et al.*, 2011] F. Zhuang, P. Luo, H. Xiong, Q. He, Y. Xiong, and Z. Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. *SADM*, 4(1):100–114, 2011.