Multi-View Clustering via Joint Nonnegative Matrix Factorization

Jialu Liu[1] Chi Wang[1] Jing Gao[2] Jiawei Han[1]

[1]University of Illinois at Urbana-Champaign

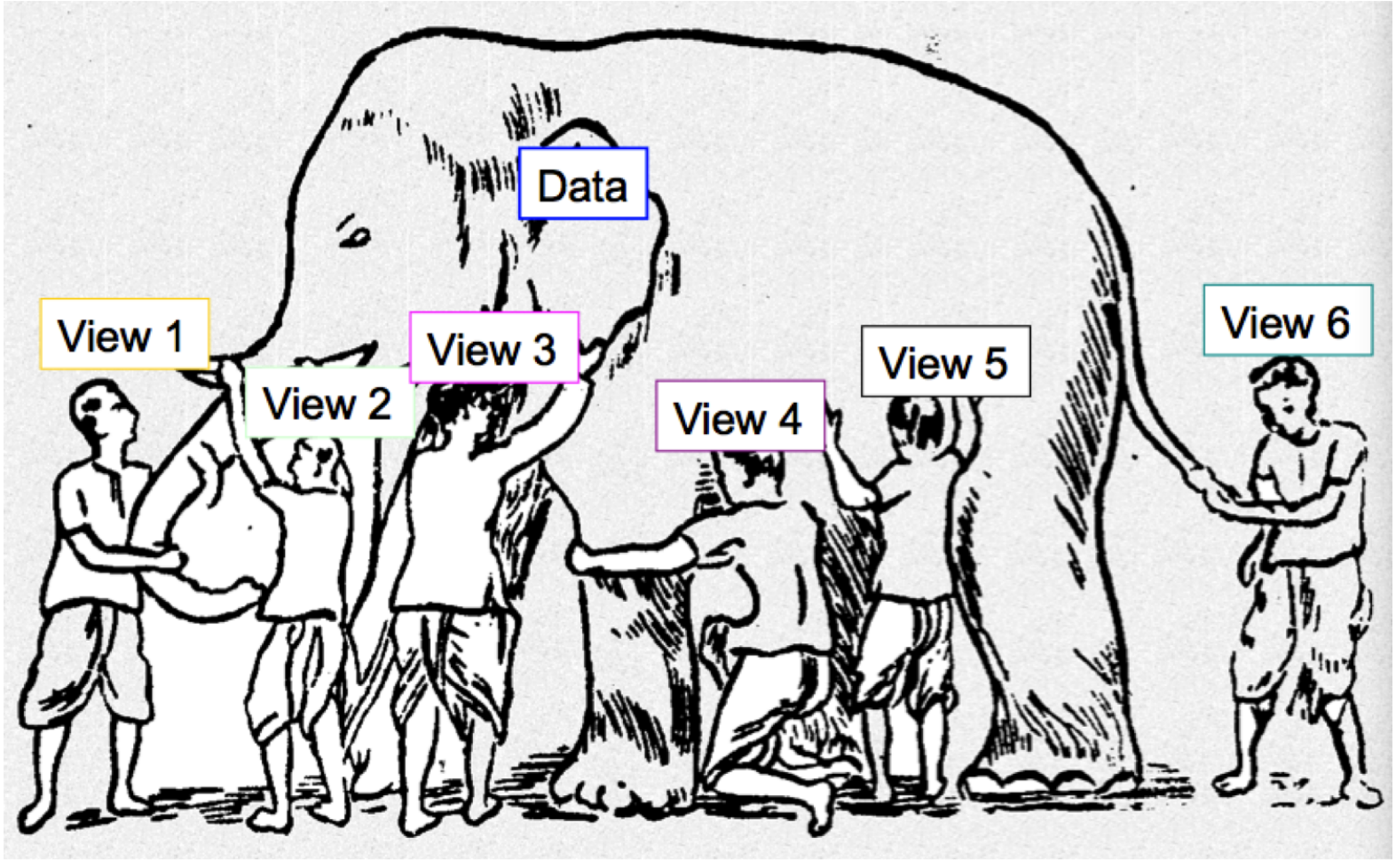
[2]University at Buffalo

May 2, 2013

Outline

Multi-View Datasets

Many datasets in real world are naturally comprised of different representations or *views*.



We need to integrate them

Observing that these multiple representations often provide *compatible* and *complementary* information, it becomes natural for one to integrate them together to obtain better performance rather than relying on a single view.

The key of learning from multiple views (multi-view) is to leverage each view's own knowledge base in order to outperform simply concatenating views.

Three ways to integrate

As we are interested in clustering, here are three common strategies.

1. Incorporating multi-view integration into the clustering process directly through optimizing certain loss functions.
2. First projecting multi-view data into a common lower dimensional subspace and then applying any clustering algorithm such as k-means to learn the partition.
3. Late integration or late fusion, in which a clustering solution is derived from each individual view and then all the solutions are fused base on consensus

Outline

Nonnegative Matrix Factorization

Let $X = [X_{\cdot,1}, \ldots, X_{\cdot,N}] \in \mathbb{R}_+^{M \times N}$ denote the nonnegative data matrix where each column represents a data point and each row represents one attribute. NMF aims to find two non-negative matrix factors $U = [U_{i,k}] \in \mathbb{R}_+^{M \times K}$ and $V = [V_{j,k}] \in \mathbb{R}_+^{N \times K}$ whose product provides a good approximation to X:

$$X \approx UV^T \tag{1}$$

Here K denotes the desired reduced dimension, and to facilitate discussions, we call U the *basis matrix* and V the *coefficient matrix*.

Update Rule of NMF

One of the common reconstruction processes can be formulated as a *Frobenius norm* optimization problem, defined as:

$$\min_{U,V} ||X - UV^T||_F^2, \ s.t. \ U \geq 0, V \geq 0$$

Multiplicative update rules are executed iteratively to minimize the objective function as follows:

$$U_{i,k} \leftarrow U_{i,k} \frac{(XV)_{i,k}}{(UV^TV)_{i,k}}, \ V_{j,k} \leftarrow V_{j,k} \frac{(X^TU)_{j,k}}{(VU^TU)_{j,k}} \tag{2}$$

NMF for Clustering

Note that given the NMF formulation in Equation 1, for arbitrary invertible $K \times K$ matrix Q, we have

$$UV^T = (UQ^{-1})(QV^T) \tag{3}$$

There can be many possible solutions, and it is important to enforce constraints to ensure uniqueness of the factorization for clustering.

One of the common ways is to normalize basis matrix U after convergence of multiplicative updates if we use V for clustering:

$$U_{i,k} \leftarrow \frac{U_{i,k}}{\sqrt{\sum_i U_{i,k}^2}}, \; V_{j,k} \leftarrow V_{j,k} \sqrt{\sum_i U_{i,k}^2} \tag{4}$$

Outline

Multi-View Notations

Assume that we are now given n_v representations (i.e., views). Let $\{X^{(1)}, X^{(2)}, \ldots, X^{(n_v)}\}$ denote the data of all the views, where for each view $X^{(v)}$, we have factorizations that $X^{(v)} \approx U^{(v)}(V^{(v)})^T$.

Here for different views, we have the same number of data points but allow for different number of attributes, hence $V^{(v)}$s are of the same shape but $U^{(v)}$s can differ along the row dimension across multiple views.

One Simple Baseline

Using the shared coefficient matrix but different basis matrices across views as shown below:

$$\sum_{v=1}^{n_v} \lambda_v \| X^{(v)} - U^{(v)} (V^{(*)})^T \|_F^2$$

where λ_v is the weight parameter and $V^{(*)}$ is the shared consensus. It is easy to verify that this baseline is equivalent to applying NMF directly on the concatenated features of all views.

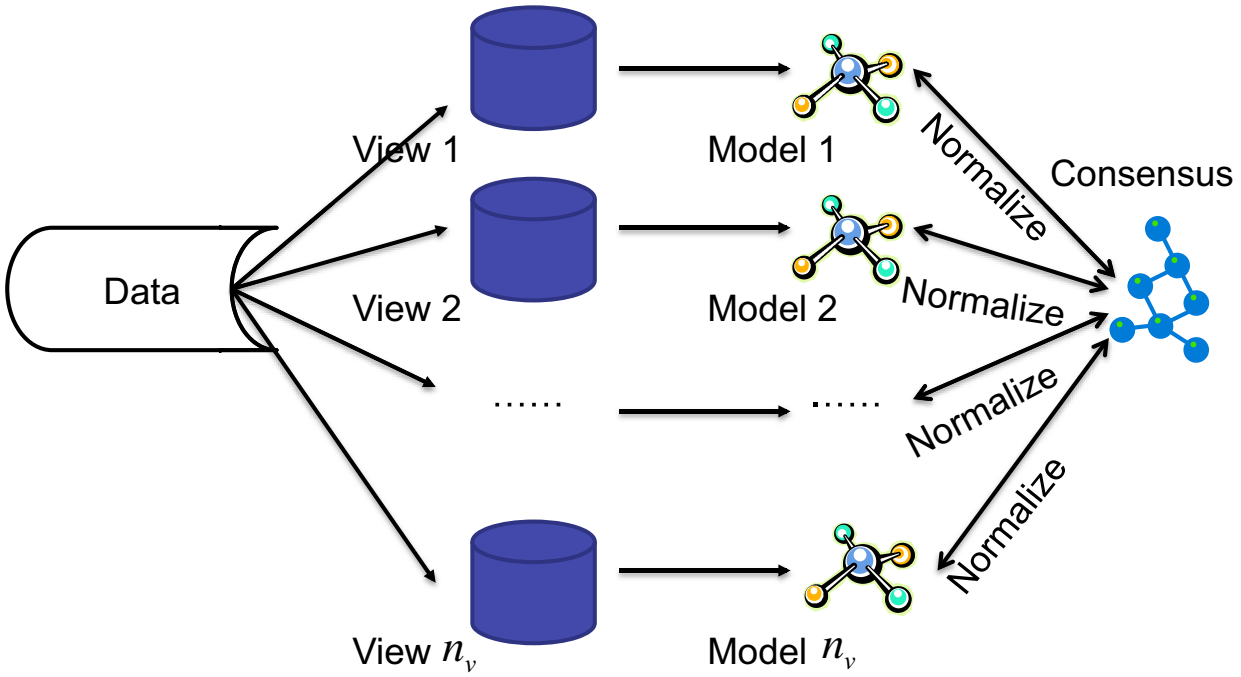
Comment

$$\sum_{v=1}^{n_v} \lambda_v \|X^{(v)} - U^{(v)}(V^{(*)})^T\|_F^2$$

However, this simple model cannot be the optimal for two reasons.

1. The hard assumption that fixing one-side factor seems to be too strong and many times we prefer relatively soft constraints.
2. With proper normalization, previous work on single-view NMF has shown to achieve better performance in terms of clustering.

Our Framework



We require models learnt from different views to be softly regularized towards a consensus with proper normalization for clustering.

Our Approach

The hard assumption that fixing one-side factor seems to be too strong and many times we prefer relatively soft constraints.

Firstly, we incorporate the disagreement between coefficient matrix $V^{(v)}$ and the consensus matrix V^* into NMF:

$$\sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$
$$s.t.\ U^{(v)}, V^{(v)}, V^* \geq 0 \tag{5}$$

Our Approach

With proper normalization, previous work on single-view NMF has shown to achieve better performance in terms of clustering.

Secondly, we add constraints on coefficient matrices $U^{(v)}$ in different views to make $V^{(v)}$s comparable and meaningful for clustering.

W.l.o.g., assume $||X^{(v)}||_1 = 1$, we then want to minimize:

$$\sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$
$$s.t.\ \forall 1 \leq k \leq K, ||U^{(v)}_{\cdot,k}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \geq 0 \qquad (6)$$

Why $||X||_1 = 1$ and $||U_{\cdot,k}||_1 = 1$?

Objective function:

$$\min_{U^{(v)}, V^{(v)}, V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$

$$s.t.\ \forall 1 \le k \le K, ||U_{\cdot,k}^{(v)}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \ge 0$$

Given $||X||_1 = 1$ and $||U_{\cdot,k}||_1 = 1$,

$$||X||_1 = ||\sum_j X_j||_1 \approx \sum_{k=1}^{K} ||U_{\cdot,k} \sum_j V_{j,k}||_1 = \sum_{k=1}^{K} ||\sum_j V_{j,k}||_1 = ||V||_1$$

Therefore,

$$||V||_1 \approx 1$$

Objective Function

Previous:

$$\min_{U^{(v)},V^{(v)},V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$

$$s.t.\ \forall 1 \le k \le K, ||U^{(v)}_{\cdot,k}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \ge 0$$

Now:

$$\min_{U^{(v)},V^{(v)},V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(Q^{(v)})^{-1}Q^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)}Q^{(v)} - V^*\|_F^2$$

$$s.t.\ \forall 1 \le v \le n_v, U^{(v)} \ge 0, V^{(v)} \ge 0, V^* \ge 0$$

where

$$Q^{(v)} = Diag\left(\sum_{i=1}^{M} U^{(v)}_{i,1}, \sum_{i=1}^{M} U^{(v)}_{i,2}, \ldots, \sum_{i=1}^{M} U^{(v)}_{i,K}\right)$$

Iterative Update Rules

Fixing V^*, minimize over $U^{(v)}$ and $V^{(v)}$ until convergence:

$$U_{i,k} \leftarrow U_{i,k} \frac{(XV)_{i,k} + \lambda_v \sum_{j=1}^{N} V_{j,k} V^*_{j,k}}{(UV^T V)_{i,k} + \lambda_v \sum_{l=1}^{M} U_{l,k} \sum_{j=1}^{N} V^2_{j,k}}$$

$$U \leftarrow UQ^{-1}, \; V \leftarrow VQ$$

$$V_{j,k} \leftarrow V_{j,k} \frac{(X^T U)_{j,k} + \lambda_v V^*_{j,k}}{(VU^T U)_{j,k} + \lambda_v V_{j,k}}$$

Fixing $U^{(v)}$ and $V^{(v)}$, minimize over V^*:

$$V^* = \frac{\sum_{v=1}^{n_v} \lambda_v V^{(v)} Q^{(v)}}{\sum_{v=1}^{n_v} \lambda_v} \geq 0$$

Use V^* for Clustering

Once we obtain the consensus matrix V^*, the cluster label of data point j could be computed as $\arg\max_k V^*_{j,k}$.

Or we can simply use k-means directly on V^* where V^* is viewed as a latent representation of the original data points.

Outline

PLSA

Probabilistic Latent Semantic Analysis (PLSA) is a traditional topic modeling technique for document analysis. It models the $M \times N$ term-document co-occurrence matrix X (each entry X_{ij} is the number of occurrences of word w_i in document d_j) as being generated from a mixture model with K components:

$$P(w, d) = \sum_{k=1}^{K} P(w|k)P(d, k)$$

Relation to NMF

$$P(w,d) = \sum_{k=1}^{K} P(w|k)P(d,k)$$

$$X = (UQ^{-1})(QV^T)$$

Early studies show that (UQ^{-1}) (or (QV^T)) has the formal properties of conditional probability matrix $[P(w|k)] \in \mathbb{R}_+^{M\times K}$ (or $[P(d,k)]^T \in \mathbb{R}_+^{K\times N}$). This provides theoretical foundation for using NMF to conduct clustering.

Due to this connection, joint NMF has a nice probabilistic interpretation: each element in the matrix V^* is the consensus of $P(d|k)^{(v)}$ weighted by $\lambda_v P(d)^{(v)}$ from different views.

Outline

Datasets

One synthetic and three real world datasets are used in the experiment.

- **Synthetic dataset:** It is a two-view dataset where noises are added independently.
- **3-Sources Text dataset:** It is collected from three online news sources: BBC, Reuters, and The Guardian telling the same story.
- **Reuters Multilingual dataset:** This test collection contains feature characteristics of documents originally written in different languages.
- **UCI Handwritten Digit dataset:** This handwritten digits (0-9) data is from the UCI repository with different features.

Datasets

The important statistics of four datasets are summarized in the following table.

dataset	size	# view	# cluster
Synthetic	10000	2	4
3-Sources	169	3	6
Reuters	600	3	6
Digit	2000	2	10

Compared Algorithms

We compared with the following algorithms:

- Single View (**BSV** and **WSV**): Runing each view using the NMF technique. Then both the best and the worst single view results are reported, which are referred to as BSV and WSV respectively.
- Feature Concatenation (**ConcatNMF**): Concatenating the features of all the views, and then run NMF directly on this concatenated view representation. The normalization strategy is adopted.
- Collective NMF (**CoINMF**): Using the shared coefficient matrix but different basis matrices across views.
- Co-regularized Spectral clustering (**Co-reguSC**): Adopting co-regularization framework to spectral clustering.
- Multi-View NMF (**MultiNMF**): This is the proposed algorithm. In our experiments, we empirically set λ_v to 0.01 for all views and datasets.

Performance

Algorithm	Accuracy(%)			
	Synthetic	3-Sources	Reuters	Digit
BSV	66.0±.09	60.8±.01	46.8±.02	68.5±.05
WSV	51.7±.11	49.1±.03	46.4±.00	63.4±.04
ConcatNMF	68.4±.14	58.6±.03	47.3±.00	67.8±.06
ColNMF	61.8±.08	61.3±.02	51.2±.00	66.0±.05
Co-reguSC	75.4±.00	47.8±.01	50.6±.02	86.6±.00
MultiNMF	**92.0±.10**	**68.4±.06**	**53.5±.00**	**88.1±.01**

Algorithm	Normalized Mutual Information(%)			
	Synthetic	3-Sources	Reuters	Digit
BSV	56.2±.10	53.0±.01	38.8±.02	63.4±.03
WSV	54.3±.05	44.1±.02	34.2±.00	60.3±.03
ConcatNMF	60.9±.14	51.7±.03	34.1±.00	62.4±.04
ColNMF	47.3±.07	55.2±.02	34.6±.00	62.1±.03
Co-reguSC	71.2±.00	41.4±.01	35.7±.01	77.0±.00
MultiNMF	**84.0±.15**	**60.2±.06**	**40.9±.00**	**80.4±.01**

The higher, the better for both *Accuracy* and *Normalized Mutual Information*.

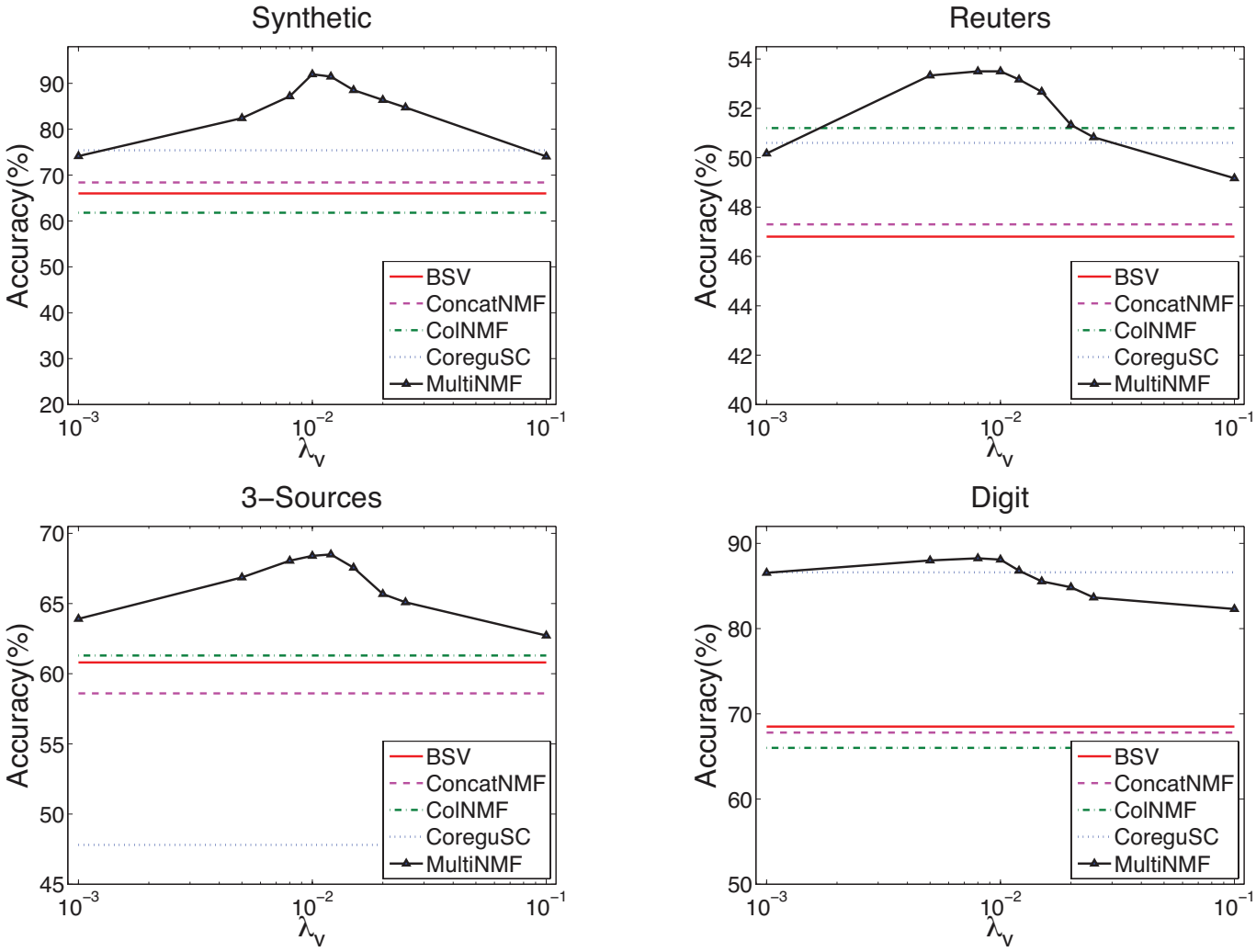
Parameter Study

There are n_v parameters in our MultiNMF algorithm: the regularization parameters λ_v for each view.

- The relative value of λ_v among multiple views reflects each view's importance.
- The absolute value of λ_v reflects how much we want to enforce the regularization constraint.
- In the extreme case, when λ_vs are all 0, the problem reduces to the same as doing NMF with normalization for each view seperately; when λ_vs go to infinity, $V^{(v)}Q^{(v)}$ for different views share the same value.

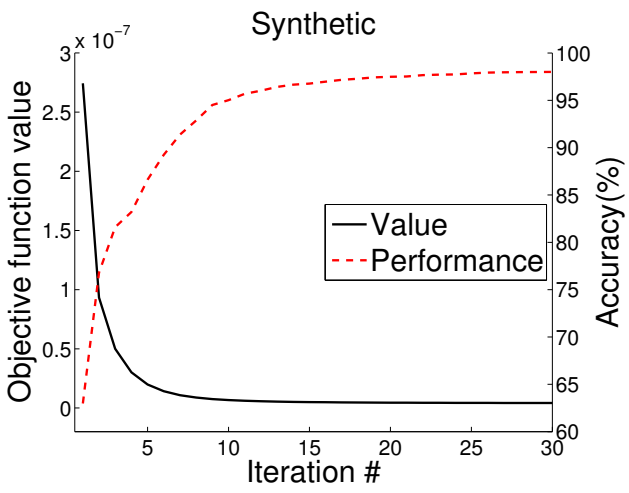
Parameter Study

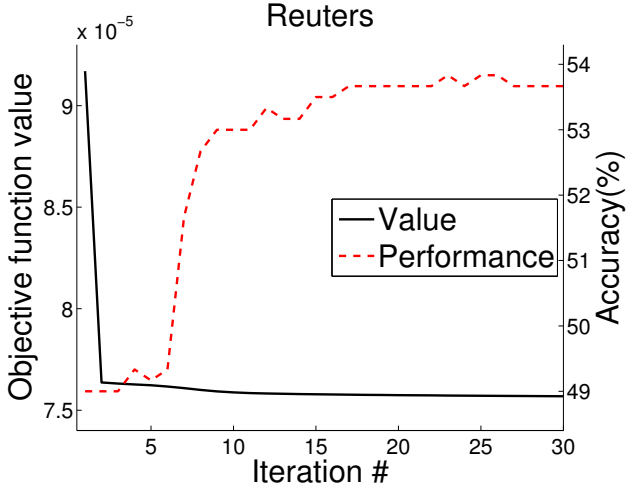
We study the absolute value of λ_v here.

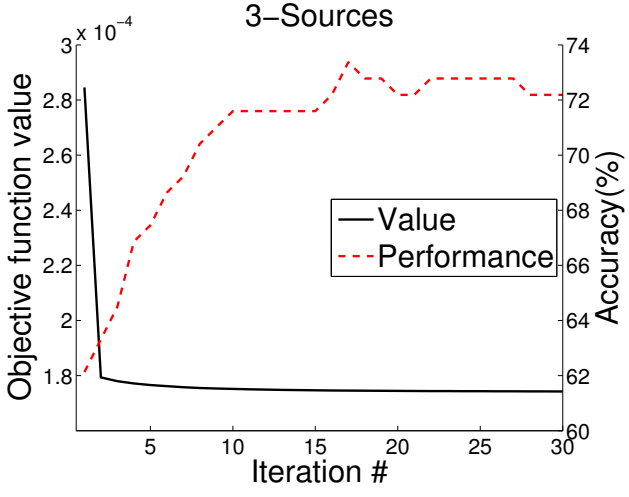


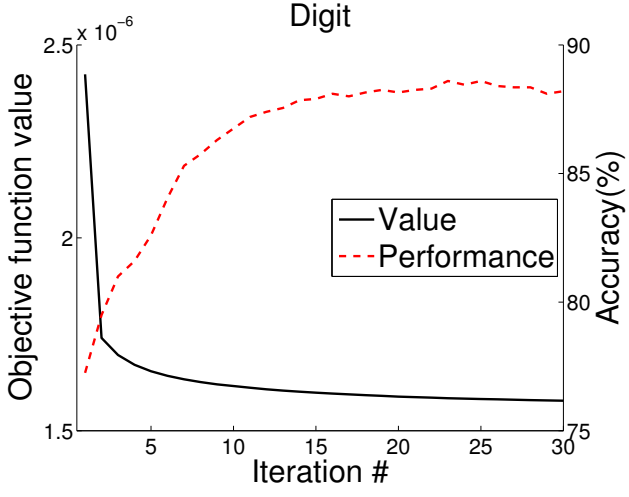
Convergence Study

We prove that the multiplicative update rules are convergent in the paper. Figure below shows the convergence curve together with its performance.



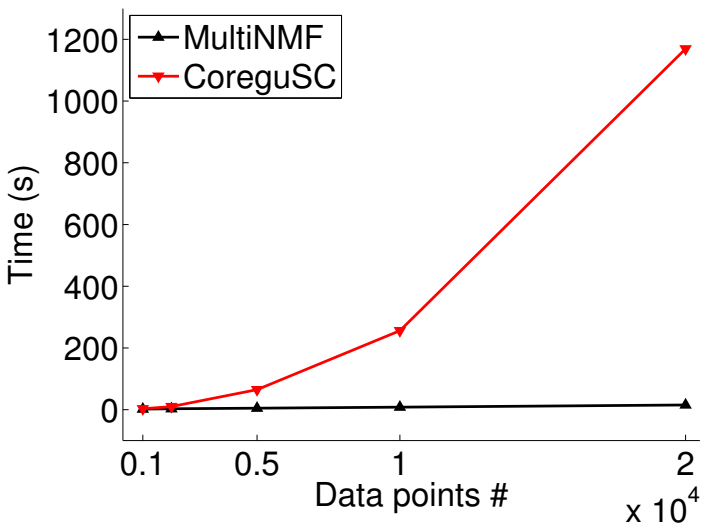


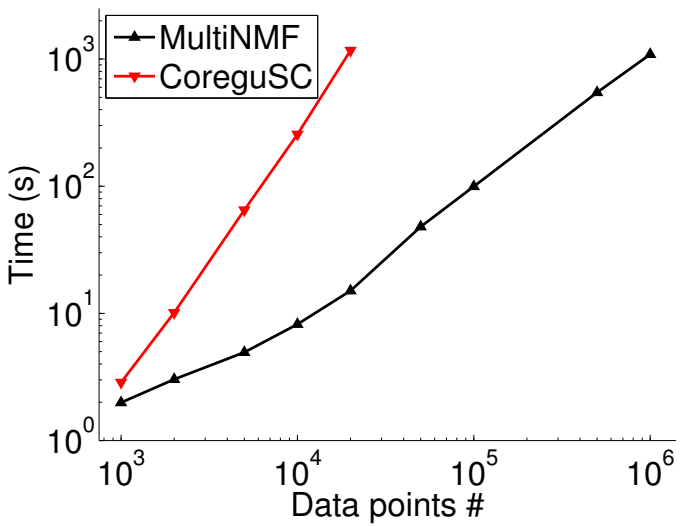




Computational Complexity Study

MultiNMF has linear time complexity in the number of data points, clusters, and views. We conduct experiments on the synthetic dataset. The default setting is 10000 data points, 4 clusters, and 2 views. During the experiment, we fix two aspects and change the remaining one.





Computational Complexity Study