PARTIAL MULTI VIEW CLUSTERING

VIA NON-NEGATIVE MATRIX FACTORIZATION

- Introduction
- Existing Methods
- Multi View Clustering
- Problem Variants
- Proposed Models
- Results
- Future Work

16th December, 2015
Xerox Research Centre India

Introduction

Motivation :

- Given multiple views (i.e. sets of features) for an instance. how to utilize the information in all the views to get better results.
- Information in different views might complement each other and may give better results if you appropriately.

Challenges :

- Different views
- No obvious way to compare different views
- Merge information while respecting every view

Existing Work

1. Multi View K Means : Extending K means to multiple views. Modify cost function to scale to multiple views.
2. Spectral Clustering : Multiple approaches, related to Bipartite graphs, Reconstruction of similarity matrix, Co regularization of Clustering Hypothesis, etc
3. Non Negative Matrix Factorization : Topic of today's discussion
4. Many More

Non Negative Matrix Factorization

Advantages:

- Due to the nonnegativity constraints, the NMF produces a so-called “additive parts-based” representation of the data. Consequently, the factors W and H are generally naturally sparse.
- Impressive benefits in terms of interpretation of its factors

Formulation:

Let X (m x 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 (m x k) and V (k x n), whose product provides a good approximation to X i.e. X = UV’ (‘ : Transpose)

Computation:

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

Multiplicative Update

Multi View Clustering Via NMF

Formulation:

Assume that we are now given n_v representations (i.e. views). Let { X_1, X_2 . . . X_{n_v} } denote the data of all the views, where for each view X_v , we have factorizations that $X_v \approx U_v V_v^T$. In standard NMF, coefficient vector V_j can be regarded as low-rank representation of the jth data point in terms of the new basis U_v

Objective function:

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

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

Challenge:

Ensuring that V_v's remain comparable across views. Add constraint of U for the same reason.

Variants To Multi View Clustering

Many variants to classic Multi View setup. Aimed at modelling real world data more realistically. The famous variants are as follows,

1. **Partial Multi View Clustering (PVC) :** A much more realistic model. Not all instances have complete views (i.e. do not have all sets of views). The setup remains roughly the same but with existence of partial views for instances.
2. Constrained Multi View Clustering on Unmapped data : Let's make the problem more difficult, we have multiple views of instances. But don't know which view belongs to which instance (i.e. unmapped). Existence of inter-view constraints which help in aiding us.

Existing Work on PVC

Let us consider the simple case of two views, Without considering inter view information, we simply get the loss function shown in (A).

But to use information from both views, we modify the function, or more correctly the variables involved as shown in (B). We keep a common part across views for the complete instances (i.e. the one with both views)

(Xc are instances with complete views)

A

$$\min_{U^{(1)}\geq 0,\bar{P}^{(1)}\geq 0} \|\bar{\mathbf{X}}^{(1)} - \bar{P}^{(1)}U^{(1)}\|_F^2 + \lambda\Omega(\bar{P}^{(1)}),$$

$$\min_{U^{(2)}\geq 0,\bar{P}^{(2)}\geq 0} \|\bar{\mathbf{X}}^{(2)} - \bar{P}^{(2)}U^{(2)}\|_F^2 + \lambda\Omega(\bar{P}^{(2)}),$$

B

$$\begin{aligned} \min_{\{U^{(v)},\bar{P}^{(v)}\}_{v=1}^{2}} O \equiv & \left\| \begin{bmatrix} \mathbf{X}_c^{(1)} \\ \hat{\mathbf{X}}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_F^2 + \lambda\|\bar{P}^{(1)}\|_1 \\ + & \left\| \begin{bmatrix} \mathbf{X}_c^{(2)} \\ \hat{\mathbf{X}}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_F^2 + \lambda\|\bar{P}^{(2)}\|_1 \\ \text{s.t.} \quad & U^{(1)} \geq 0, U^{(2)} \geq 0, \\ & \bar{P}^{(1)} \geq 0, \bar{P}^{(2)} \geq 0, \end{aligned} \tag{3}$$

Proposed Models

There are three major models we have proposed and compared,

1. Graph Regularization with Hard Constraints
2. Graph Regularization with Soft Constraints
3. Model (2) with varying weights

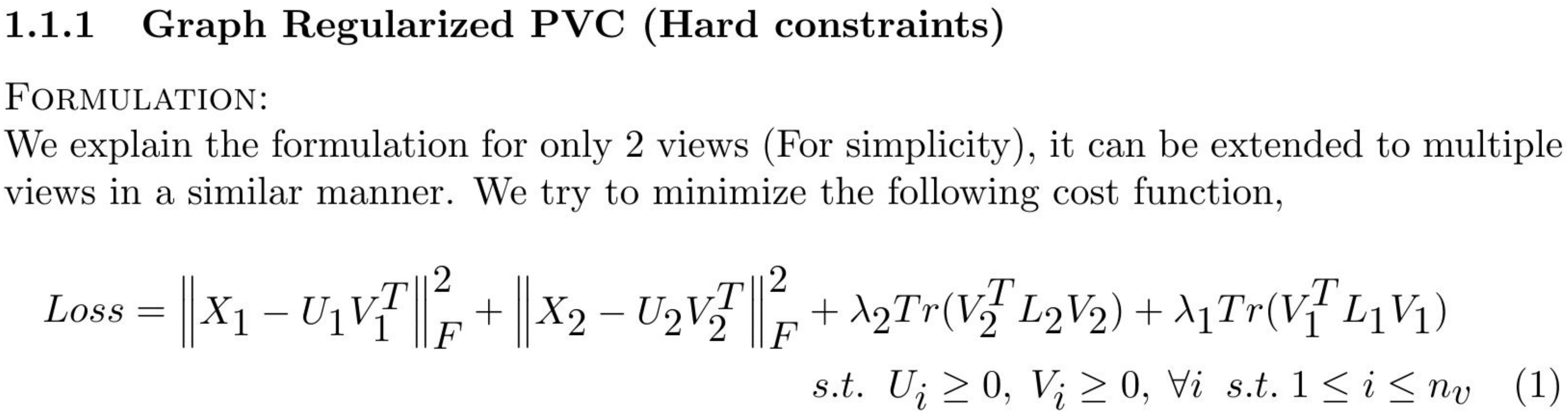
Graph Regularization (Basics)

Motivation:

Main aim is to reduce the distance (in the final latent space) between the instances according to the original space. We construct a matrix W, which may roughly represent the closeness (similarity) between two points. The final objective becomes,

$$\begin{aligned}\mathcal{R}_1 &= \frac{1}{2}\sum_{j,l=1}^{N} \|\mathbf{z}_j - \mathbf{z}_l\|^2 \mathbf{W}_{jl} \\ &= \sum_{j=1}^{N} \mathbf{z}_j^T \mathbf{z}_j \mathbf{D}_{jj} - \sum_{j,l=1}^{N} \mathbf{z}_j^T \mathbf{z}_l \mathbf{W}_{jl} \\ &= \mathrm{Tr}(\mathbf{V}^T\mathbf{D}\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{W}\mathbf{V}) = \mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}),\end{aligned}$$

1. Graph Regularized PVC



1.1.1 Graph Regularized PVC (Hard constraints)

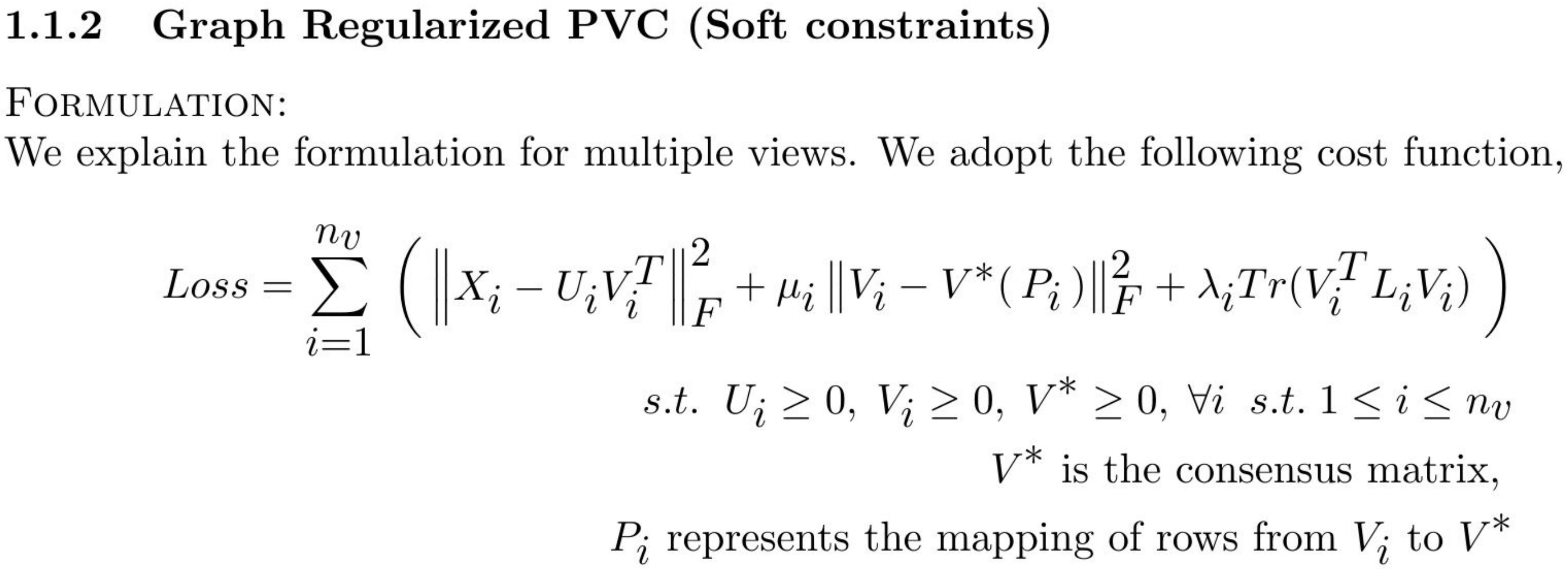
Formulation:
We explain the formulation for only 2 views (For simplicity), it can be extended to multiple views in a similar manner. We try to minimize the following cost function,

$$Loss = \left\|X_1 - U_1V_1^T\right\|_F^2 + \left\|X_2 - U_2V_2^T\right\|_F^2 + \lambda_2 Tr(V_2^T L_2 V_2) + \lambda_1 Tr(V_1^T L_1 V_1)$$

$$s.t.\ \ U_i \geq 0,\ V_i \geq 0,\ \forall i\ \ s.t.\ 1 \leq i \leq n_v \qquad (1)$$

Based on the PVC approach discussed earlier

2. Graph Regularized PVC



1.1.2 Graph Regularized PVC (Soft constraints)

FORMULATION:
We explain the formulation for multiple views. We adopt the following cost function,

$$Loss = \sum_{i=1}^{nv} \left(\left\|X_i - U_i V_i^T\right\|_F^2 + \mu_i \left\|V_i - V^*(P_i)\right\|_F^2 + \lambda_i Tr(V_i^T L_i V_i) \right)$$

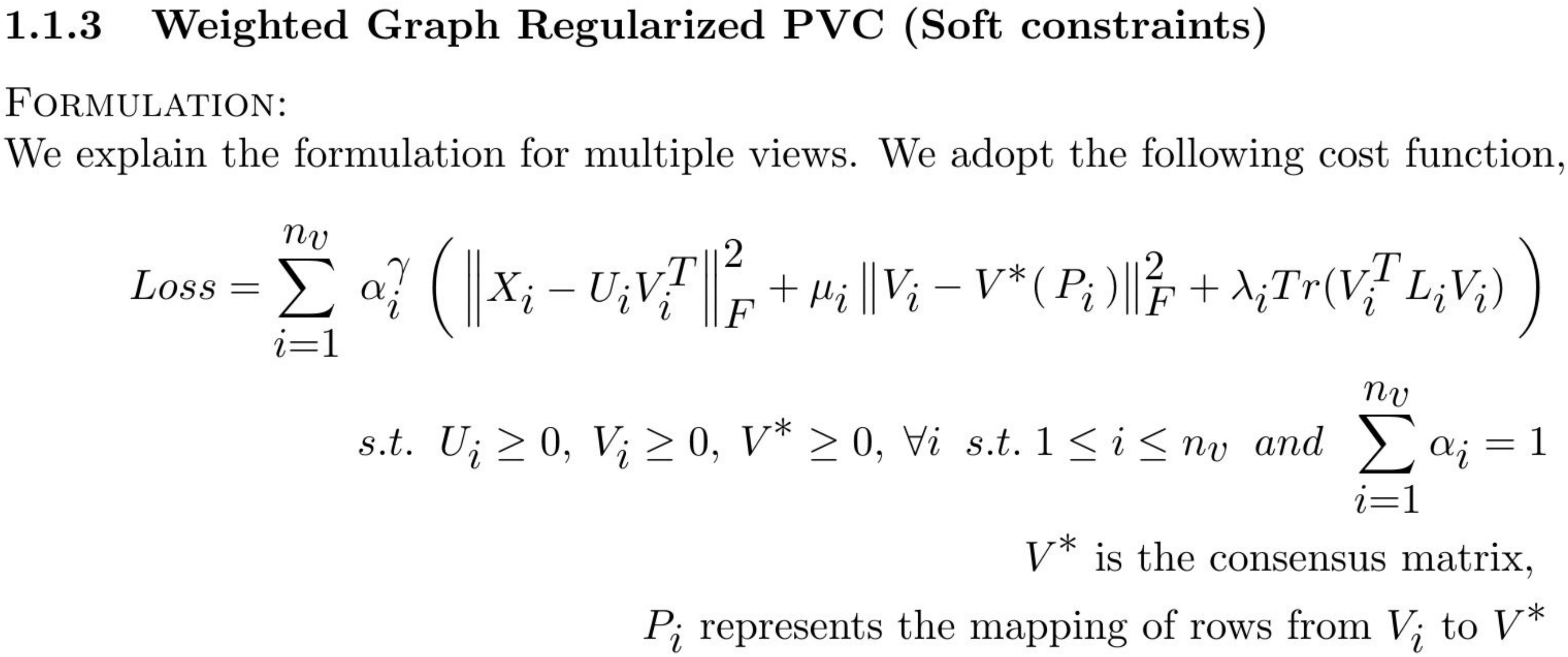
$$s.t.\ \ U_i \geq 0,\ V_i \geq 0,\ V^* \geq 0,\ \forall i\ \ s.t.\ 1 \leq i \leq nv$$

V^* is the consensus matrix,

P_i represents the mapping of rows from V_i to V^*

Based on the Multi View approach discussed earlier

3. Weighted Graph Regularized PVC



1.1.3 Weighted Graph Regularized PVC (Soft constraints)

FORMULATION:
We explain the formulation for multiple views. We adopt the following cost function,

$$Loss = \sum_{i=1}^{nv} \alpha_i^{\gamma} \left(\left\| X_i - U_i V_i^T \right\|_F^2 + \mu_i \left\| V_i - V^*(P_i) \right\|_F^2 + \lambda_i Tr(V_i^T L_i V_i) \right)$$

$$s.t.\ \ U_i \geq 0,\ V_i \geq 0,\ V^* \geq 0,\ \forall i\ \ s.t.\ 1 \leq i \leq nv\ \ and\ \sum_{i=1}^{nv} \alpha_i = 1$$

V^* is the consensus matrix,

P_i represents the mapping of rows from V_i to V^*

Results

Datasets used,

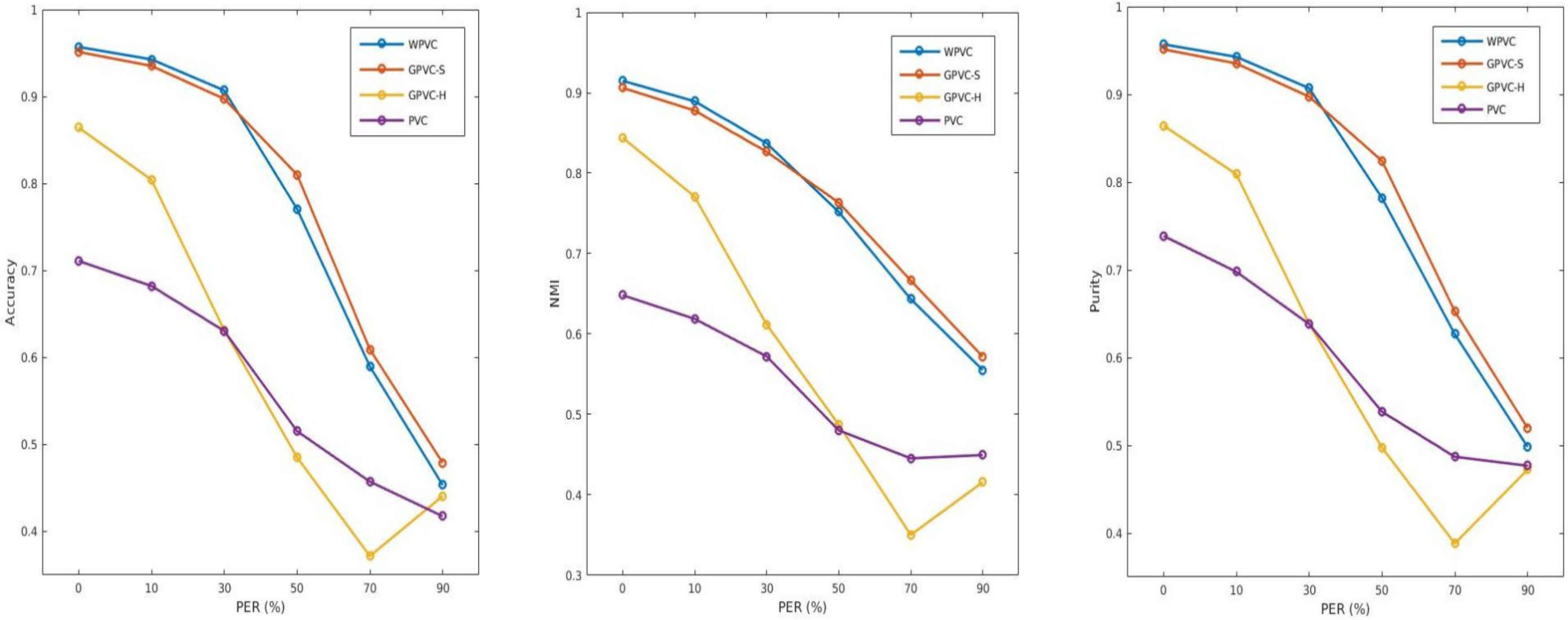
- **UCI Handwritten Digit dataset :** This hand-written digits (0-9) data is from the UCI repository. The dataset consists of 2000 examples, with view-1 being the 76 Fourier coefficients and view-2 being the 240 pixel averages in 2x3 windows.
- **ORL :** Collection of facial images of 40 subjects. Consists of 400 images with 2 views each.
- **BBCSport :** Collection of sports news reports from BBC. Consists of 544 reports with 2 views each

Comparison Measures

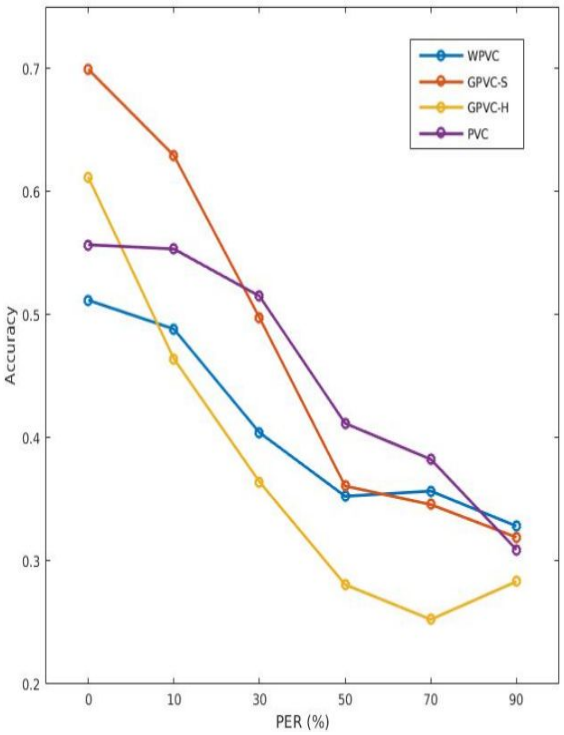
Measures used for comparison,

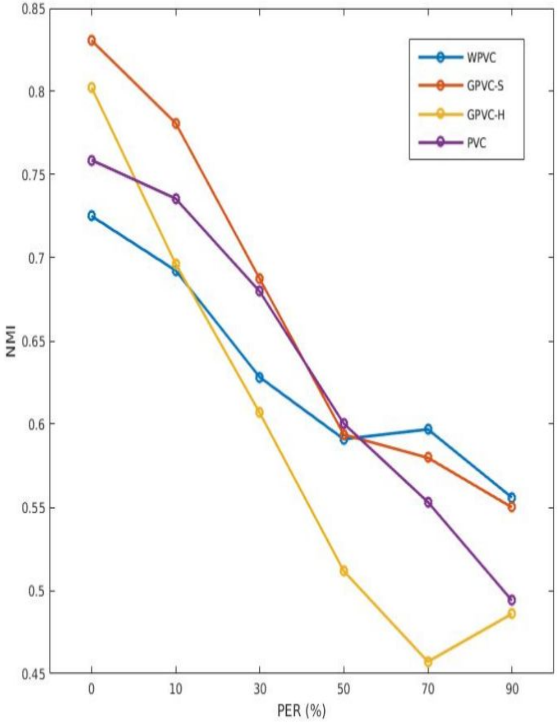
- **Accuracy :** There is no obvious way of computing accuracy in case of clustering. Instead, we find the best map of the clusters to the classes (Using Hungarian Algorithm) and compute the measure.
- **Normalized Mutual Information :** Measure of how much knowing one thing can tell you about the other thing.
- **Purity :** Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by the total samples i.e. $\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$

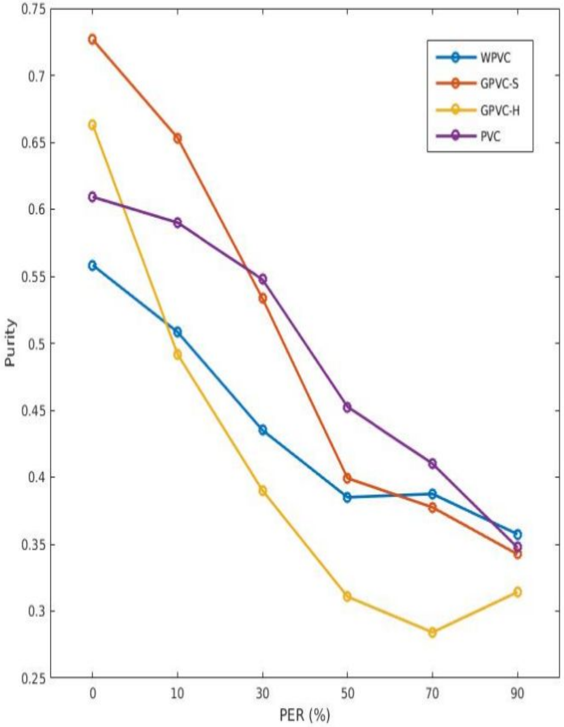
Results - Digit



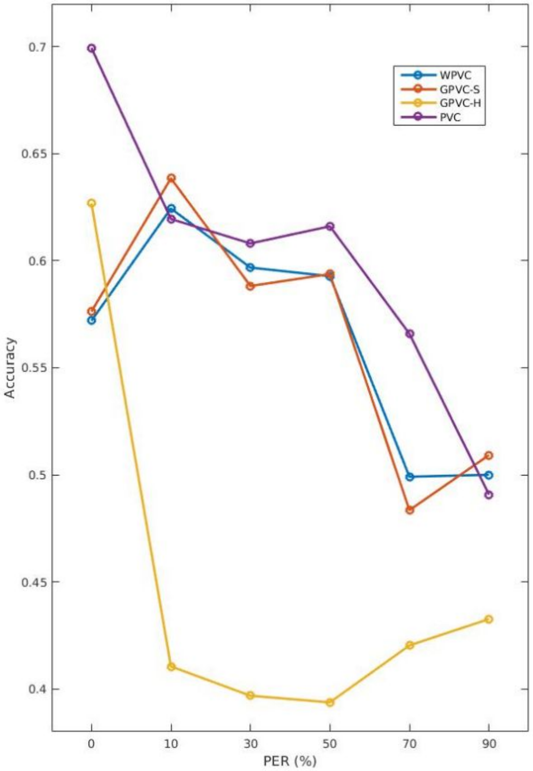
Results - ORL Database

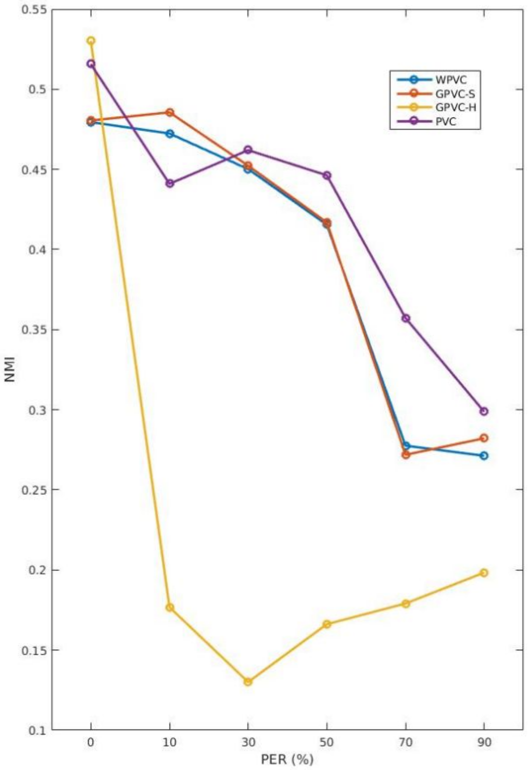


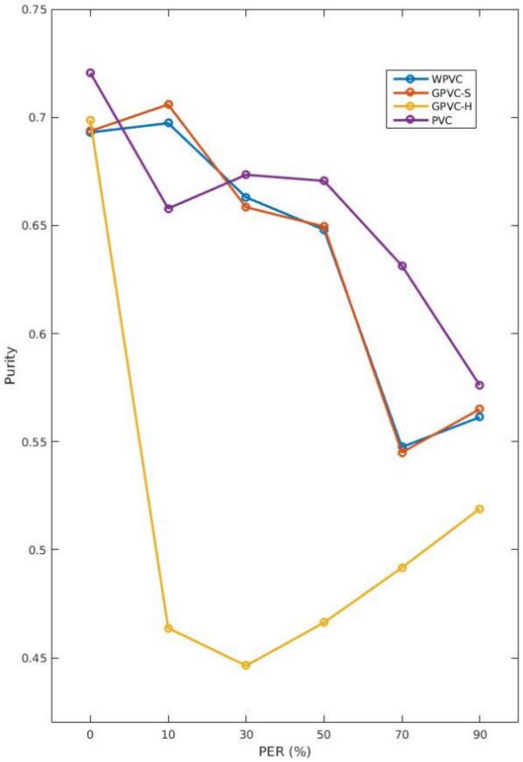




Results - BBCSports







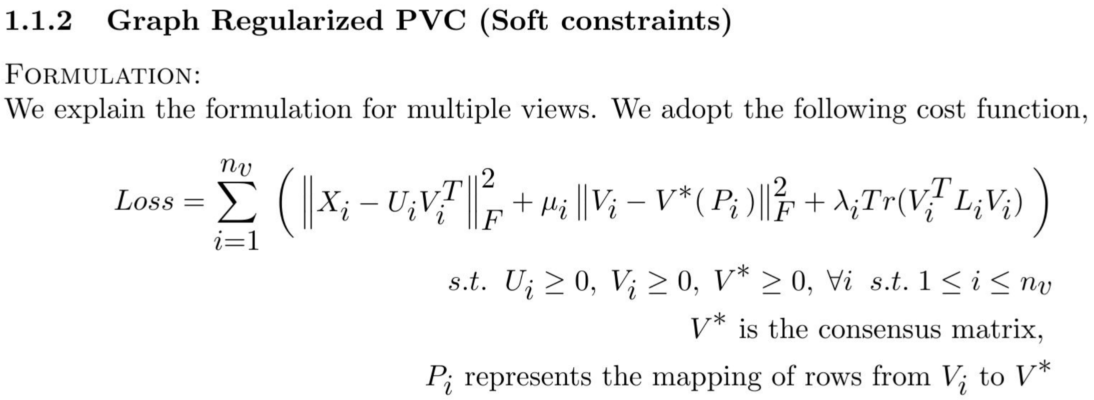
Observations

Conclusions from the previous results,

- Perform **better** (much better) than PVC (previous work) in case of **Image** datasets (i.e. Digit, ORL)
- Performance of GPVC-Hard **degrades** very fast.
- **Poor** results in case of **text** datasets. Suspect better results due to the **lasso** norm included in the setup of PVC (Lasso works great for text).
- **Hard** constraints seem to work well with BBCSport dataset (Text datasets).
- Possible issue with the optimization **algorithm** itself. PVC uses **Greedy Coordinate Descent** while we use **Multiplicative Updates**.

Extension to K views

- No existing work with clear description of extending the partial view setup to K views.
- Main idea is to use the maximum amount of information possible. So, if an instance has t views present, try to use all the t views to get the final result.
- Extending the previous setup of GPVC-S to k dimensions. The formulation remains the same as discussed earlier.



1.1.2 Graph Regularized PVC (Soft constraints)

FORMULATION:
We explain the formulation for multiple views. We adopt the following cost function,

$$Loss = \sum_{i=1}^{n_v} \left(\left\| X_i - U_i V_i^T \right\|_F^2 + \mu_i \left\| V_i - V^*(P_i) \right\|_F^2 + \lambda_i Tr(V_i^T L_i V_i) \right)$$

$$s.t.\ \ U_i \geq 0,\ V_i \geq 0,\ V^* \geq 0,\ \forall i\ \ s.t.\ 1 \leq i \leq n_v$$

V^* is the consensus matrix,

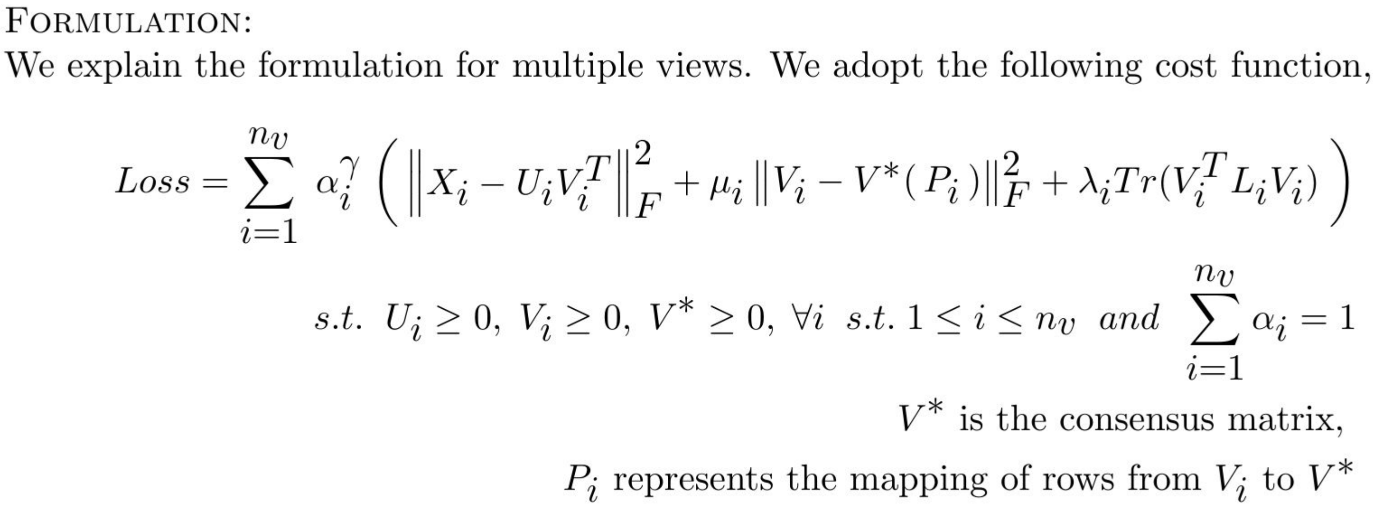
P_i represents the mapping of rows from V_i to V^*

Experiments:
- Artificially create Partial View data.
- Random process
- Vary the complete view ratio
- Distribute the remaining instances equally amongst the other views.
- Not realistic but gives a rough idea of how the system performs in presence of partial view data

Weighted Graph Regularized Multi View Clustering

Formulation similar to the one for the Partial View Clustering.
In fact, Partial view clustering is a general version of Multi View clustering.

RECAP



FORMULATION:
We explain the formulation for multiple views. We adopt the following cost function,

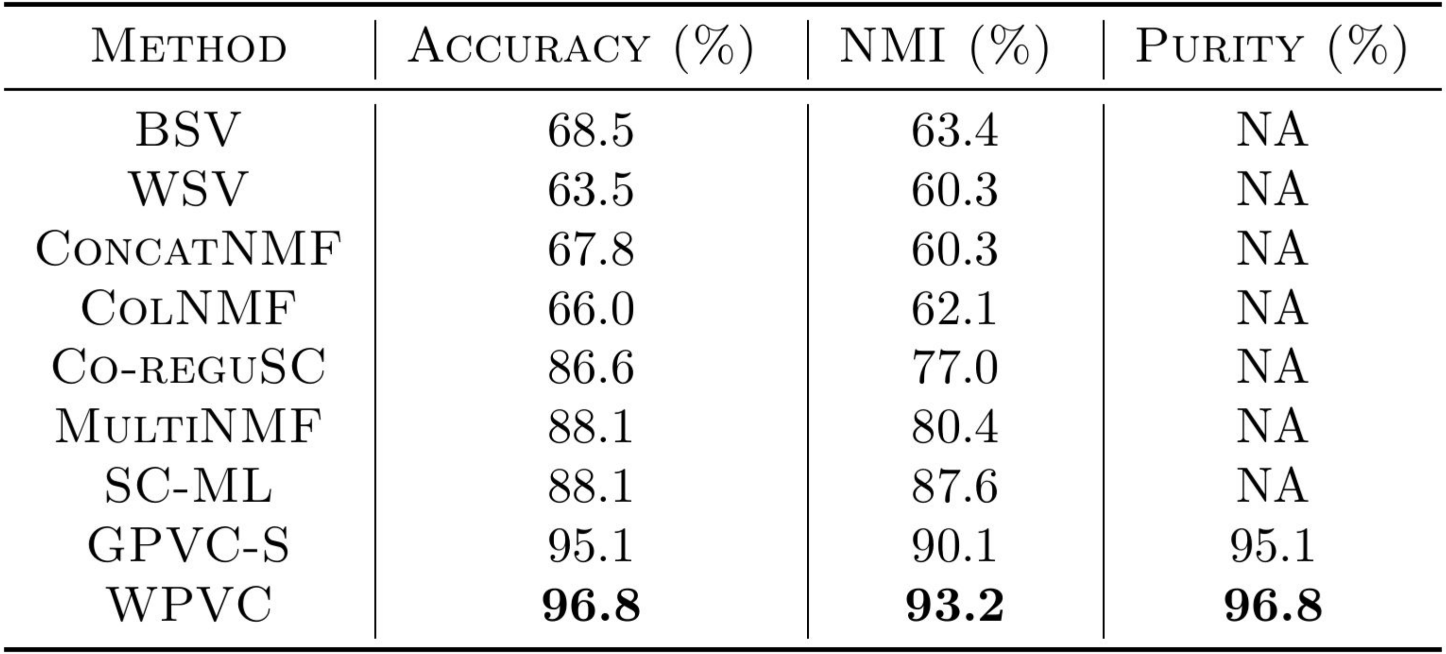
$$Loss = \sum_{i=1}^{nv} \alpha_i^{\gamma} \left(\left\| X_i - U_i V_i^T \right\|_F^2 + \mu_i \left\| V_i - V^*(\,P_i\,) \right\|_F^2 + \lambda_i Tr(V_i^T L_i V_i) \right)$$

$$s.t.\ \ U_i \geq 0,\ V_i \geq 0,\ V^* \geq 0,\ \forall i\ \ s.t.\ 1 \leq i \leq nv\ \ and\ \sum_{i=1}^{nv} \alpha_i = 1$$

V^* is the consensus matrix,

P_i represents the mapping of rows from V_i to V^*

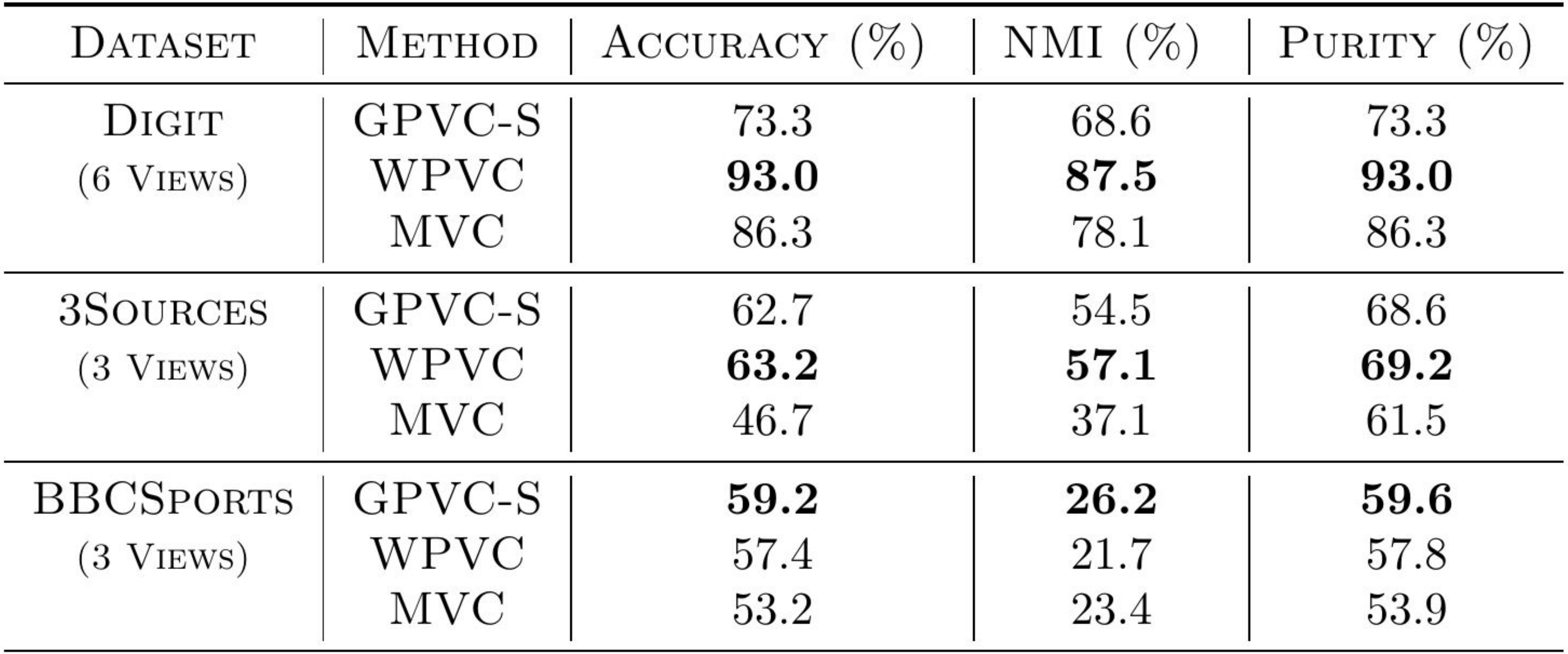
Results - Digit (2 Views)



Method	Accuracy (%)	NMI (%)	Purity (%)
BSV	68.5	63.4	NA
WSV	63.5	60.3	NA
ConcatNMF	67.8	60.3	NA
ColNMF	66.0	62.1	NA
Co-reguSC	86.6	77.0	NA
MultiNMF	88.1	80.4	NA
SC-ML	88.1	87.6	NA
GPVC-S	95.1	90.1	95.1
WPVC	**96.8**	**93.2**	**96.8**

Results for Digit (2 views)

Results (Continued)



Dataset	Method	Accuracy (%)	NMI (%)	Purity (%)
Digit (6 Views)	GPVC-S	73.3	68.6	73.3
	WPVC	**93.0**	**87.5**	**93.0**
	MVC	86.3	78.1	86.3
3Sources (3 Views)	GPVC-S	62.7	54.5	68.6
	WPVC	**63.2**	**57.1**	**69.2**
	MVC	46.7	37.1	61.5
BBCSports (3 Views)	GPVC-S	**59.2**	**26.2**	**59.6**
	WPVC	57.4	21.7	57.8
	MVC	53.2	23.4	53.9

Future Work

- Incorporating **View specific feature weights** : Further stretching the idea of view wise weights to also incorporate varying feature weights for different views. Seems important due to the fact that not all features are equally important.
- Introducing **L1** normalization : Possibly improve results on **text** datasets, since it promotes **sparseness**.
- Introduction of **L21 norm** : It has been shown to be more **robust to outliers** and handles **noisy data** well.