Constrained Inference for Multi-View Clustering

Shalmali Joshi SHALMALI@UTEXAS.EDU
Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78703

Oluwasanmi Koyejo SANMI.K@UTEXAS.EDU
Imaging Research Center, University of Texas, Austin, TX 78703

Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78703

Abstract

We propose a novel approach for probabilistic multi-view clustering that combines view-specific models to improve global coherence. Global incoherence is measured by the difference between view-specific cluster assignment responsibilities. New cluster responsibilities are estimated by optimizing a cost function that maximizes per-view accuracy subject to a user-specified global coherence threshold. When combined with a parameter estimation step, this modified inference encourages the estimation of model parameters that agree between views. We show that the modified inference remains convex when global coherence constraints are given by the norm of the difference between the responsibilities of each model. In addition, the global correction is embarrassingly parallel between examples. The proposed approach is evaluated on a synthetic dataset as well as real data showing improved performance as compared to strong baseline methods for multi-view clustering.

1. Introduction

Several real world datasets consist of examples where the descriptive features can be divided into coherent subsets, also known as views. Each view generally consists of features that are statistically independent across views, conditioned on the identity of the example. An example is a scientific dataset consisting of sets of descriptive features associated with each individual such as genetic features, brain imaging features and behavioral phenotypes that have been gathered and processed independently. Another example is a tagged image dataset where each example may be described either by the raw image features, or by the tag annotations.

Multi-view learning is a popular paradigm for constructing coherent models that combine information from multiple views of the same problem. Multi-view clustering is a special case of unsupervised multi-view learning where each view is a clustering model. Specifically, probabilistic multi-view clustering introduces latent variables indicating cluster assignment for each data view. We propose a novel multiview clustering model that uses a separate cluster indicator for each view. This is combined with a global coherence constraint. Global incoherence is determined as the difference between view-specific cluster assignment responsibilities. Thus, new cluster responsibilities are estimated by optimizing a cost function that maximizes per-view accuracy subject to a user-specified global coherence threshold. When combined with a parameter estimation step, this modified inference encourages the estimation of model parameters that agree between views.

As the proposed approach requires only a global pooling and correction of per-view responsibilities, it is straightforward to apply to existing view-specific probabilistic clustering models when the existing models are trained via expectation maximization. The expectation and maximization steps remain unchanged and may be computed in parallel for each view. Further, any favorable properties of the per-view inference and maximization steps are retained. For example, the expectation step for computing cluster responsibilities is generally embarrassingly parallel across examples and the solution is often computable in closed form. We show that the global pooling step remains convex when the global coherence constraints are given by the norm of the difference between the responsibilities of each model. In addition, the global correction is embarrassingly parallel between examples.

Appearing in *Proceedings of the 31^{st} International Conference on Machine Learning*, Beijing, China, 2014.

Our main contributions are as follows:

- We propose a novel framework for multi-view clustering using constrained inference to enforce global coherence on view-specific clustering models.
- We show that the global pooling step remains convex when the global coherence constraints are given by the norm of the difference between the responsibilities of each model.
- We show improved performance on simulated data and real data as compared to strong baseline models.

The rest of the paper is organized as follows. We discuss previous work in Section 2 and discuss background material on constrained inference in Section 3. We propose the constrained inference approach for multiview clustering in Section 4, including constraints on the norm of the posterior responsibilities. Section 5 includes experiments and results.

1.1. Preliminaries

Vectors are denoted by lower-case bold $\mathbf{x}$ and matrices by capital $\mathbf{X}$. The ℓ_q-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ with $q > 0$ is given by:

$$\|\mathbf{x}\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{1}{q}}.$$

Let $Z \sim P(Z)$ be a random variable with associated probability density function $p(\mathbf{z})$. We use $\mathrm{E}_p[\,\beta(\mathbf{z})\,] = \int_{\mathbf{z}} \beta(\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$ to denote the expectation of the function $\beta(\mathbf{z})$ with respect to $Z \sim P(Z)$. The *relative entropy* (also known as the KL divergence) between two densities $p(\mathbf{z})$ and $q(\mathbf{z})$ is given by,

$$\mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) = \mathrm{E}_q[\,\log q(\mathbf{z}) - \log p(\mathbf{z})\,]. \quad (1)$$

The relative entropy is non-negative and strictly convex with respect to its first argument. Further $\mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) = 0$ iff. $q(\mathbf{z}) = p(\mathbf{z})$.

An exponential family is a class of densities parameterized by $\boldsymbol{\theta}$ of the form:

$$p(\mathbf{x}) = h(\mathbf{x}) e^{\langle \eta(\boldsymbol{\theta}), \beta(\mathbf{x}) \rangle - G(\boldsymbol{\theta})},$$

where $h(\mathbf{x})$ is the base measure, $\eta(\boldsymbol{\theta})$ is the natural parameter, $\beta(\mathbf{x})$ is a vector of sufficient statistics and $G(\boldsymbol{\theta})$ is called the log-partition function which ensures that the density integrates to one.

The *categorical distribution* is a discrete distribution over choices $y \in \{1, 2, \ldots, k\}$ parameterized by $\boldsymbol{\theta} \in \{\mathbf{a} \in \mathbb{R}^k,\ a_i \geq 0,\ \sum_i a_i = 1\}$ so that $p(Y = i) = \theta_i$. The categorical distribution is the exponential family with natural parameters $\log \boldsymbol{\theta}$ and sufficient statistics given by the vector of indicator functions for each choice, denoted by $\mathbf{z}(y) \in \mathbb{R}^K$ with:

$$z_k = \begin{cases} 1 & \text{if } y = i, \\ 0, & \text{otherwise.} \end{cases}$$

To simplify the presentation, we will follow the convention of describing the categorial distribution directly as a distribution over the indicator variables.

A vector $\mathbf{x} \in \mathbb{R}^{K+1}$ lies in the K-simplex, denoted by Δ_K, if $x_i \geq 0,\ \sum_i x_i = 1$.

2. Related Work

While multi-view learning was initially motivated for supervised learning problems (Blum & Mitchell, 1998), analogous procedures have also been proposed for unsupervised learning (Bickel & Scheffer, 2005). Multi-view learning procedures combine view-specific models via a collaborative learning procedure that trades off the quality of each view-specific model with global coherence constraints. Multi-view learning procedures are motivated by theoretical performance analysis for the supervised case (Blum & Mitchell, 1998) and the unsupervised clustering case (Bickel & Scheffer, 2005). In addition, multi-view learning has been shown to outperform direct feature concatenation in many practical scenarios (Chaudhuri et al., 2009; Cai et al., 2013).

Co-EM (Bickel & Scheffer, 2004) is a popular model for probabilistic multi-view clustering. As with naïve feature concatenation, Co-EM uses a single cluster indicator variable. However, this is combined with parameter regularization that encourages global coherence between views. Non-probabilistic clustering models for multi-view data have been addressed via spectral clustering (Zhou & Burges, 2007; Kumar et al., 2011) and canonical correlation analysis (Chaudhuri et al., 2009). Cai et al. (2013) proposed an approach to improve the scalability of multi-view clustering. Further, Bickel & Scheffer (2004) presented an extensive study of hierarchical agglomerative clustering across multiple views and a study of conventional k-means and EM based probabilistic clustering for multiview data.

Multiview clustering is also related to approaches such as cluster ensembles (Strehl & Ghosh, 2003) in which each clustering model of the ensemble is developed based on different feature sets. Co-clustering (Banerjee et al., 2007) can also be viewed as a special type of multi-view clustering where each row or column defines a view. Finally, non-parametric Bayesian approaches have also been proposed for multiview models (Li & Shafto, 2011).

3. Background: Constrained Bayesian Inference

Constrained Bayesian inference is a special case of probabilistic inference via relative entropy minimization subject to data constraints on the observed variables and additional constraints on the latent variables. The discussion in this section follows the discussion in (Koyejo & Ghosh, 2013a). Let Z be a latent variable whose density we wish to recover subject to constraints, and let $\mathbf{x}$ be the observed data. In the absence of latent variable constraints, the probabilistic inference results in the posterior density that can be computed using Bayes rule as:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}.$$

Suppose, in addition to observed data, inference includes additional latent variable constraints $\mathrm{E}\left[\,\beta(\mathbf{z})\,\right] \in \mathsf{C}$, given feature functions $\beta(\mathbf{z})$ and a constraint set C. We now wish to recover a density $q(\mathbf{z})$ that satisfies such constraints. This target density $q(\mathbf{z})$, called the post-data density (Koyejo & Ghosh, 2013b) is estimated so the result is as close as possible to the unconstrained posterior $p(\mathbf{z}|\mathbf{x})$ in terms of relative entropy. This is given as the solution of the following optimization problem (Koyejo & Ghosh, 2013a):

$$\min_{q(\mathbf{z})\in\mathcal{P},\, \mathrm{E}_q\left[\,\beta(\mathbf{z})\,\right]\in\mathsf{C}} \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})), \tag{2}$$

or equivalently, by the solution of:

$$\min_{q(\mathbf{z})\in\mathcal{P},\, \mathrm{E}_q\left[\,\beta(\mathbf{z})\,\right]\in\mathsf{C}} \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))) - \mathrm{E}_q\left[\,\log p(\mathbf{x}|\mathbf{z})\,\right]. \tag{3}$$

The inference problem (2) is typically solved using the Fenchel or Lagrange dual approach (Altun & Smola, 2006). However, Koyejo & Ghosh (2013a) proposed an alternative representation approach with two main components. The first step is identification of the parametric family of the solution, and the second step is optimization directly over that parametric family. The following proposition describes the parametric family of the feasible solution.

Proposition 1 (Koyejo & Ghosh (2013a)). *For any feasible point* $\mathbf{c} \in \mathsf{C}$*, the minimizer of:*

$$\min_{q(\mathbf{z})\in\mathcal{P},\, \mathrm{E}_q\left[\,\beta(\mathbf{z})\,\right]=\mathbf{c}} \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) \tag{4}$$

is given by $q_{\mathbf{c}}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})e^{\langle \boldsymbol{\lambda}_{\mathbf{c}}, \boldsymbol{\beta}(\mathbf{z})\rangle - G(\boldsymbol{\lambda}_{\mathbf{c}})}$.

Koyejo & Ghosh (2013a) show that as a result, any solution of (2) takes the parametric form of a member of the exponential family $f_{\boldsymbol{\theta}}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})e^{\langle \boldsymbol{\lambda}(\boldsymbol{\theta}), \boldsymbol{\beta}(\mathbf{z})\rangle - G(\boldsymbol{\lambda}(\boldsymbol{\theta}))}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the associated mean parameter, and $\boldsymbol{\Theta}$ is the mean parameter space. Further, it is shown that the solution can be found as the optimizer of the relative entropy minimization problem directly over members of this family as:

$$\min_{\mathbf{c}\in\mathsf{C}} \left[\min_{q(\mathbf{z})\in\mathcal{P},\mathrm{E}_q\left[\,\beta(\mathbf{z})\,\right]=\mathbf{c}} \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) \right],$$

The result is the following optimization approach for constrained relative entropy inference.

Corollary 2 (Koyejo & Ghosh (2013a)). *The minimizer of* (2) *is given by* $q_* = f_{\boldsymbol{\theta}_*}$ *where* $\boldsymbol{\theta}_*$ *is the solution of:*

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\begin{array}{l} \mathrm{KL}(f_{\boldsymbol{\theta}}(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) \\ s.t.\ \mathrm{E}_{f_{\boldsymbol{\theta}}}\left[\,\boldsymbol{\beta}(\mathbf{z})\,\right] \in \mathsf{C} \end{array} \right].$$

Other applications of constrained Bayesian inference include maximum entropy discrimination Maximum Entropy Discrimination (Jaakkola et al., 1999). More recently, researchers have applied constrained Bayesian inference for combining nonparametric topic models with support vector machines inspired by large margin constraints for document classification (Zhu et al., 2009), multitask classification (Zhu et al., 2011) and link prediction (Zhu, 2012).

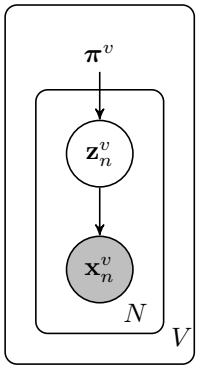

Figure 1. Generative Model for Multi-view Clustering. Each view is denoted by v

4. A Constrained Inference Approach for Multi-view Clustering

We propose a novel constrained inference approach for multi-view clustering. Let N be the total number of samples and V represent the total number of views. Each view uses the same number of clusters K. Let $n \in [N]$, $v \in [V]$ and $k \in [K]$ index the sample index, view index and cluster index respectively. The cluster membership for each example in each view is denoted by $\mathbf{z}_n^v$, an indicator variable for the choice of cluster index k. Further, the feature variable corresponding to each sample index n and view v is denoted by $\mathbf{x}_n^v$. Given a prior cluster indicator distribution for

each view $\boldsymbol{\pi}^v$, the generative model is given by:

$$Z_n^v \sim P(Z_n^v; \boldsymbol{\pi}^v),\ X_n^v \sim P(X_n^v | \mathbf{z}_n^v, \boldsymbol{\pi}^v)\, \forall\, n \in [N], v \in [V],$$

with corresponding densities $p(\mathbf{z}_n^v; \boldsymbol{\pi}^v)$ for the prior and $p(\mathbf{x}_n^v | \mathbf{z}_n^v, \boldsymbol{\pi}^v)$ for the likelihood respectively.

Without additional constraints, the above probabilistic clustering model is independent for each view, and the posterior density can be found in closed form using Bayes rule (or approximate inference methods for more challenging problems). We therefore assume that an estimate of the posterior $p(\mathbf{z}_v | \mathbf{x}_v, \boldsymbol{\pi}^v)$ is tractable to compute and is available. The posterior distribution is a categorical distribution. Without loss of generality, this estimate is parameterized by $\boldsymbol{\theta}_n^v \in \mathbb{R}^K$. Next, we introduce global coherence constraints that ensure the predicted cluster indices agree with each other on average. This is given by the constraint:

$$E_q[Z_n^v - Z_n^w] \in \mathsf{C}\ \ \forall\, v \in [V], w \in [V], n \in [N],$$

for $q > 0$ and some constraint set C. Next we combine these constraints with the posterior distribution $p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\pi}^v)$ using the constrained inference approach outlined in (2).

Applying the results of Corollary 2, the post-data density $q(\{\mathbf{z}_n^v\}_{v\in[V]} | \mathbf{x}_n^v, \boldsymbol{\pi}_v)$ takes the form $q(\{\mathbf{z}_n^v\}_{v\in[V]}) \propto$:

$$\left(\prod_{v\in[V]} p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\pi}^v)\right)\left(\prod_{u,w\in[V]\times[V]} e^{\left\langle \boldsymbol{\lambda}_{(u,w)}, \mathbf{z}_n^u - \mathbf{z}_n^w \right\rangle}\right),$$

and re-arranging terms, we find that:

$$q(\{\mathbf{z}_n^v\}_{v\in[V]}) \propto \prod_{v\in[V]}\left(p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\pi}^v) \prod_{w\in[V]} e^{\left\langle \boldsymbol{\lambda}_{(v,w)}, \mathbf{z}_n^v - \mathbf{z}_n^w \right\rangle}\right).$$

Henceforth, we will drop the conditioning on $\mathbf{x}$ and π^v for convenience of notation and denote the unconstrained posterior $p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\pi}^v)$ as $p(\mathbf{z}_n^v)$. The post-data density is independent between views as $q(\{\mathbf{z}_n^v\}_{v\in[V]}) = \prod_{v\in[V]} q(\mathbf{z}_n^v)$, with shared parameters $\boldsymbol{\lambda}_{(v,w)}$. This independence implies that the constrained inference takes the form:

$$\begin{aligned} &\min_{\{q(\mathbf{z}_n^v)\}} \sum_{v\in[V]} \mathrm{KL}(q(\mathbf{z}_n^v) \| p(\mathbf{z}_n^v)) \\ &\text{s.t. } E_q[Z_n^v - Z_n^w] \in \mathsf{C}\ \forall v \in [V], w \in [V], n \in [N], \end{aligned} \tag{5}$$

4.1. Multi-view Clustering using Norm Constraints

For the remainder of this manuscript we focus on the special case where the global coherence is enforced by norm constraints of the form:

$$\sum_{v\in[V],w\in[V]} \|\mathrm{E}_q\left[\, Z_n^v - Z_n^w \,\right]\|_p^r \le \gamma\ \forall\, n \in [N],\ p > 1, r \ge 1. \tag{6}$$

Convexity of (6) is sufficient to imply the uniqueness of constrained relative entropy minimizer (5) (Koyejo & Ghosh, 2013a). The following proposition shows that (6) is indeed a convex set.

Proposition 3. *Let*

$$\mathsf{S} = \{\{\mathbf{b}^v \in \mathbb{R}^K\} \,|\, \sum_{v,w\in[V]\times[V]} \|\mathbf{b}^v - \mathbf{b}^w\|_p^r \le \gamma,\ p > 1, r \ge 1\},$$

then S *is a convex set.*

Proof. First, we prove convexity of the ℓ_p-norm of the pair-wise difference between views. Consider the convex combination of two members of this set, given by $\mathbf{c}^v = \alpha\mathbf{a}^v + (1-\alpha)\mathbf{b}^v$, with $\alpha \in (0,1)$, then by applying standard properties of norms, we have that:

$$\begin{aligned} &\sum_{v\in[V],w\in[V]} \|\mathbf{c}^v - \mathbf{c}^w\|_p \\ &= \sum_{v\in[V],w\in[V]} \|\alpha(\mathbf{a}^v - \mathbf{a}^w) + (1-\alpha)(\mathbf{b}^v - \mathbf{b}^w)\|_p \\ &\le \sum_{v\in[V],w\in[V]} \|\alpha(\mathbf{a}^v - \mathbf{a}^w)\|_p + \|(1-\alpha)(\mathbf{b}^v - \mathbf{b}^w)\|_p \\ &= \sum_{v\in[V],w\in[V]} \alpha\|\mathbf{a}^v - \mathbf{a}^w\|_p + (1-\alpha)\|\mathbf{b}^v - \mathbf{b}^w\|_p. \end{aligned}$$

Define

$$\begin{aligned} h(x) &= x^r\ \forall\ x \ge 0 \\ &= 0\ \forall\ x\ <\ 0 \\ g &= \|\ \|_p \end{aligned}$$

Case 1, $K = 1$: We have that $\mathbf{dom}\, g = \mathbf{dom}\, h = \mathbb{R}$, $g(x)$ is convex and $g(x)$ and $h(x)$ are twice differentiable. Also $h(x)$ is convex and non-decreasing for $r \ge 1$. Then following the composition of convex functions from Chapter 3, (Boyd & Vandenberghe, 2004) we have that $h(g(x))$ is convex if $h(x)$ is convex, nondecreasing and g is convex. Hence S is convex for $K = 1$.

Case 2, $K > 1$: Define a function $\tilde{h}$ such that

$$\begin{aligned} \tilde{h} &= h(x)\ \forall\ x \in \mathbf{dom}\ h \\ &= \infty \text{ otherwise} \end{aligned}$$

Again using rules for composition of convex functions we have that $h(g(x))$ is convex if $h(x)$ is convex, $\tilde{h}$ is non-decreasing and $g(x)$ is convex. Hence, S is convex. □

We denote the categorical distribution parameters of the postdata distribution by $\boldsymbol{\phi}_{n,k}^v = q(\mathbf{z}_n^v = k)$. The relative

Algorithm 1 Constrained Inference for Multiview clustering

Given data $\mathbf{x}_n^v$, Initialize $\{\boldsymbol{\phi}_n^v\}, \gamma$
repeat
  for all n, v in parallel **do**
    Standard E-step to obtain $\boldsymbol{\theta}_n^v$
  end for
  for all n in parallel **do**
    Solve (7) to obtain $\{\boldsymbol{\phi}_n^v\}$
  end for
  for all v in parallel **do**
    M-step with $\boldsymbol{\theta}_n^v$ replaced by $\{\boldsymbol{\phi}_n^v\}$
  end for
until Converged

entropy cost function is given explicitly as:

$$\min_{\{\boldsymbol{\phi}_n^v\}} \sum_{v\in[V]} \langle \boldsymbol{\phi}_n^v, \log \boldsymbol{\phi}_n^v - \log \boldsymbol{\theta}_n^v \rangle$$
$$\text{s.t.} \sum_{v\in[V], w\in[V]} \|\boldsymbol{\phi}_n^v - \boldsymbol{\phi}_n^w\|_p^r \leq \gamma \; \forall \; n \in [N], \; \boldsymbol{\phi}_n^v \in \Delta_{K-1}$$

The constraint can also be converted to regularization form using Lagrange multipliers ζ_n resulting in the cost function:

$$\min_{\{\boldsymbol{\phi}_n^v\}} \sum_{v\in[V]} \langle \boldsymbol{\phi}_n^v, \log \boldsymbol{\phi}_n^v - \log \boldsymbol{\theta}_n^v \rangle + \zeta_n \sum_{v\in[V], w\in[V]} \|\boldsymbol{\phi}_n^v - \boldsymbol{\phi}_n^w\|_p^r \; \forall \; n \in [N]$$
$$\text{s.t.} \;\; \boldsymbol{\phi}_n^v \in \Delta_{K-1}$$

When combined with a parameter estimation step (M-step), this modified inference encourages the estimation of model parameters that agree between views. Algorithm 1 illustrates the combined expectation maximization algorithm with global correction step. The E-step proceeds as in the unconstrained case. The M-step also remains the same with $\{\boldsymbol{\theta}_n^v\}$ replaced by $\{\boldsymbol{\phi}_n^v\}$. Every vector $\boldsymbol{\phi}_n^v$ lies on the K-simplex, thus the optimization can be solved via projected gradient descent subject to simplex constraints.

5. Experiments

We demonstrate the effectiveness of our method using synthetic and real datasets. It seems worthwhile to verify that completely parallel E-steps and M-steps with an intermediate pooling step suffices to capture the underlying latent structure, while still allowing scalability.

Performance of unsupervised learning methods is measured using Normalized Mutual Information (Strehl et al., 2000) and Average Entropy Metric (Bickel & Scheffer, 2005). NMI quantifies the dependence between the random variables representing the true class label and the clustering result. Hence higher NMI is better. Qualitatively, Average Entropy measures the average number of bits needed to encode real class labels given the clustering result. Thus lower Average Entropy is better.

If $p_{i|j}$ is the fraction of elements of class i in cluster j and n_j be the size of cluster j. Then the average entropy metric H is given by (Bickel & Scheffer, 2005):

$$H = \sum_{j=1}^{m} \frac{n_j}{n} (-\sum_{i=1}^{C} p(i|j) \log p(i|j)) \tag{7}$$

We measure these metrics w.r.t ground-truth to compare performance with baseline models as shown in Fig. 2. In addition, specifically for the proposed model, Fig. 3 shows performance of each view in comparison with ground-truth ('View-1' and 'View-2'), the consensus performance ('Total') and a measure of the agreement between the views ('In-view'). In-view performance is an average of the Entropy measured for each view considering another view as ground-truth. Similarly for NMI.

We compare the proposed model (RAMVC - **R**epresentation **A**pproach (to Constrained Bayesian Inference) for **M**ulti-**V**iew **C**lustering) with baselines:

1. Co-EM (Bickel & Scheffer, 2005)
2. learning each view independently without constraints (Ind)
3. a single model where features from both views share the latent variable (Joint).

For the proposed model (RAMVC), the regularization parameter is set to same value for all data points as $\zeta_n = \zeta \; \forall \; n \in [N]$ and selected via cross-validation in the range $\in (10^{-15}, 10^{15})$. We used the squared ℓ_2 norm for the constraint set in all experiments. The regularization parameter of Co-EM η is cross-validated in its range (0, 1). The other baselines are unregularized.

We present experimental results with synthetic data generated from $V = 2$ views. One view corresponds to a multinomial view of the data instance and the second view has features that are Gaussian distributed. The Gaussian mixture view has dimensionality much smaller than the multinomial view. To generate a synthetic mixture model with $K = 5$ topics in 2 views, K Dirichlet distributed parameter vectors are sampled, and then used to generate the parameters of the multinomial view. Similarly for the view corresponding to Gaussian distributed features, K mean parameters were randomly generated. To generate the samples, first a topic k is chosen according to a prior probability

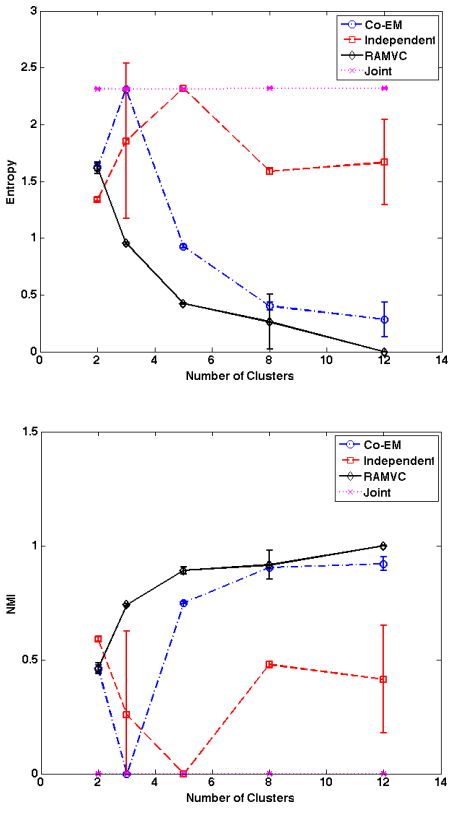

Figure 2. Average Entropy (lower is better) and NMI (higher is better) compared to baselines

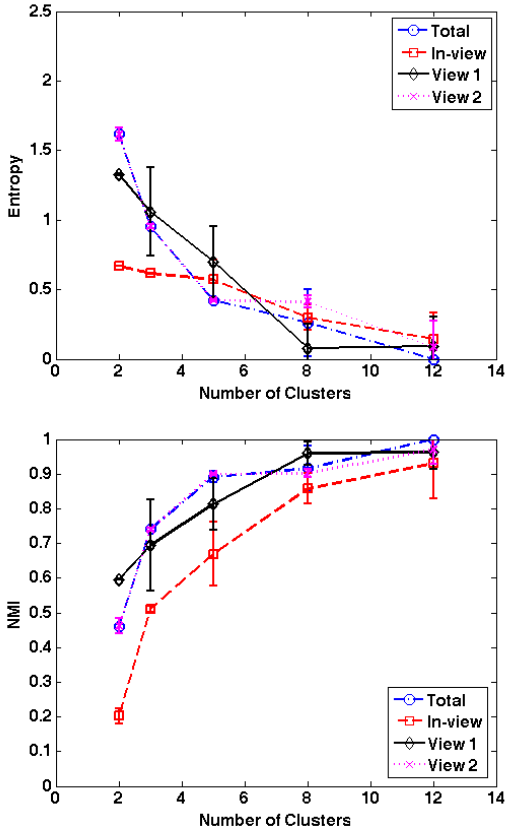

Figure 3. Average Entropy and NMI vs number of clusters - Synthetic data

π_k for every sample. Given the topic, instances for each view are sampled using the parameters chosen above for the given topic.

The generated data was analyzed using the proposed model and competing baselines. We tested the results to verify the accuracy of the cluster membership estimated by each view, and estimated by the joint model on test data. In the ideal case, we expect that both the joint cluster predictions and the individual view predictions will be accurate. Further, the individual view predictions should agree with each other. As shown in Fig. 3, the proposed model behaves as expected when between view and individual view performance is evaluated. Further, as shown in Fig. 2, RAVMC consistently outperforms all other baselines.

The proposed approach is also evaluated on the WebKB data. WebKB dataset consists of text views and link occurrences as different views. It consists of about 4500 academic web-pages where the text view corresponds to text on the webpage instance and the link views are binary vectors indicating presence of an inbound link. Approximately 100000 features are present in text view versus 3000 in the link views. The true value of K is known to be 6 but not used while training the model. The performance of the model on a hold-out dataset compared to baselines as measured w.r.t. NMI and Average Entropy is shown in Table 1 and Table 2. The performance is comparable to that of Co-EM and in general better than concatenating features or learning models across different views independently.

In addition to the clustering metrics, we evaluated the cluster prediction performance directly using classification metrics, after Hungarian matching (Kuhn, 1955) to align cluster indices with true labels. The metrics evaluated are: classification accuracy, precision, recall and f-measure. As can be seen from Table 3, the proposed method outperforms Co-EM as well as other baselines, particularly as measured by the f-measure. This further demonstrates the effectiveness of the proposed approach.

6. Conclusion

We have proposed a novel framework for probabilistic multi-view clustering using constrained inference to enforce global coherence on view-specific clustering models. The proposed approach is easily incorporated into existing

Table 1. Average Entropy vs number of clusters for WebKB data

	NUMBER OF CLUSTERS		
MODEL	2	5	8
IND	2.207(0.05)	**2.136(0.028)**	2.148(0.024)
JOINT	2.184(0.030)	2.160(0.049)	2.145(0.024)
CO-EM	2.184(0.032)	2.175(0.037)	2.137(0.034)
RAMVC	**2.181(0.059)**	2.159(0.057)	**2.128(0.013)**

Table 2. NMI vs number of clusters for WebKB data

	NUMBER OF CLUSTERS		
MODEL	2	5	8
IND	0.007(0.003)	0.019(0.003)	**0.028(0.006)**
JOINT	**0.007(0.003**	0.016(0.005)	0.015(0.009)
CO-EM	**0.007(0.004)**	0.013(0.005)	0.023(0.007)
RAMVC	**0.007(0.004)**	**0.022(0.009)**	0.025(0.008)

Table 3. Classification Accuracy using the true number of clusters ($K = 6$)

MODEL	ACC.	PRECISION	RECALL	F-MEAS.
IND	**0.2554**	0.0194	**0.5000**	0.0374
JOINT	0.2264	0.0164	0.2000	0.0302
COEM	0.2000	0.0154	0.2778	0.0292
RAMVC	0.2024	**0.1857**	**0.5000**	**0.0623**

expectation maximization (EM) methods for probabilistic clustering, requiring only the addition of a global pooling step. Further, it is shown that the global pooling step remains convex when the global incoherence constraints are given by the norm of the difference between the responsibilities of each model. For future work, we plan to explore the selection of the constraint set in more detail, to determine theoretical and empirical properties of different constraint sets in different scenarios.

References

Altun, Yasemin and Smola, Alex. Unifying divergence minimization and statistical inference via convex duality. In *Proc. of Conf. on Learning Theory (COLT*, 2006.

Banerjee, Arindam, Dhillon, Inderjit, Ghosh, Joydeep, Merugu, Srujana, and Modha, Dharmendra S. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, 2007.

Bickel, Steffen and Scheffer, Tobias. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pp. 19–26, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2142-8. URL http://dl.acm.org/citation.cfm?id=1032649.1033432.

Bickel, Steffen and Scheffer, Tobias. Estimation of mixture models using co-em. In Gama, Joo, Camacho, Rui, Brazdil, Pavel, Jorge, Alpio, and Torgo, Lus (eds.), *ECML*, volume 3720 of *Lecture Notes in Computer Science*, pp. 35–46. Springer, 2005. ISBN 3-540-29243-8. URL http://dblp.uni-trier.de/db/conf/ecml/ecml2005.html#BickelS05.

Blum, Avrim and Mitchell, Tom. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pp. 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. doi: 10.1145/279943.279962. URL http://doi.acm.org/10.1145/279943.279962.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Cai, Xiao, Nie, Feiping, and Huang, Heng. Multi-view k-means clustering on big data. In Rossi, Francesca (ed.), *IJCAI*. IJCAI/AAAI, 2013. ISBN 978-1-57735-633-2. URL http://dblp.uni-trier.de/db/conf/ijcai/ijcai2013.html#CaiNH13a.

Chaudhuri, Kamalika, Kakade, Sham M., Livescu, Karen, and Sridharan, Karthik. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 129–136, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553391. URL http://doi.acm.org/10.1145/1553374.1553391.

Jaakkola, Tommi, Meila, Marina, and Jebara, Tony. Maximum entropy discrimination. In *NIPS*, 1999.

Koyejo, Oluwasanmi and Ghosh, Joydeep. A representation approach for relative entropy minimization with expectation constraints. *International Conference on Machine Learning, The 1st Workshop on Divergences and Divergence Learning*, 2013a.

Koyejo, Oluwasanmi and Ghosh, Joydeep. Constrained bayesian inference for low rank multitask learning. *Uncertainty in Artificial Intelligence*, 2013b.

Kuhn, H.W. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97, 1955.

Kumar, Abhishek, Rai, Piyush, and III, Hal Daum. Co-regularized multi-view spectral clustering. In Shawe-Taylor, John, Zemel, Richard S., Bartlett, Peter L., Pereira, Fernando C. N., and Weinberger, Kilian Q. (eds.), *NIPS*, pp. 1413–1421, 2011. URL http://dblp.uni-trier.de/db/conf/nips/nips2011.html#KumarRD11.

Li, Dazhuo and Shafto, Patrick. Bayesian hierarchical cross-clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 443–451, 2011.

Strehl, Alexander and Ghosh, Joydeep. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

Strehl, Alexander, Ghosh, Joydeep, and Mooney, Raymond. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pp. 58–64, 2000.

Zhou, Dengyong and Burges, Christopher J. C. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 1159–1166, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273642. URL http://doi.acm.org/10.1145/1273496.1273642.

Zhu, Jun. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

Zhu, Jun, Ahmed, Amr, and Xing, Eric P. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*, 2009.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Infinite Latent SVM for Classification and Multi-task Learning. In *NIPS*, 2011.