Mach Learn (2016) 104:411–439
DOI 10.1007/s10994-016-5543-2

Rényi divergence minimization based co-regularized multiview clustering

Shalmali Joshi[1] · **Joydeep Ghosh**[1] · **Mark Reid**[2] · **Oluwasanmi Koyejo**[3]

Received: 15 March 2015 / Accepted: 4 January 2016 / Published online: 16 February 2016

Abstract Multiview clustering is a framework for grouping objects given multiple views, e.g. text and image views describing the same set of entities. This paper introduces co-regularization techniques for multiview clustering that explicitly minimize a weighted sum of divergences to impose coherence between per-view learned models. Specifically, we iteratively minimize a weighted sum of divergences between posterior memberships of clusterings, thus learning view-specific parameters that produce similar clusterings across views. We explore a flexible family of divergences, namely Rényi divergences for co-regularization. An existing method of probabilistic multiview clustering is recovered as a special case of the proposed method. Extensive empirical evaluation suggests improved performance over a variety of existing multiview clustering techniques as well as related methods developed for information fusion with multiview data.

Keywords Multiview · Clustering · Rényi divergence · Co-regularization

Editors: Concha Bielza, Joao Gama, Alipio Jorge, and Indrė Žliobaitė.

✉ Shalmali Joshi
shalmali@utexas.edu

Joydeep Ghosh
jghosh@utexas.edu

Mark Reid
mark.reid@anu.edu.au

Oluwasanmi Koyejo
sanmi@illinois.edu

[1] The University of Texas at Austin, Austin, TX, USA

[2] The Australian National University & NICTA, Canberra, ACT, Australia

[3] Stanford University & University of Illinois at Urbana-Champaign, 201 N Goodwin Ave,
Urbana, IL 61801, USA

1 Introduction

Multiple views of entities are often readily available in modern datasets, for example, a webpage entity has text, images and hyper-links, each of which can be considered as views of the web-page entity. A problem of practical interest is to harness complementary information available in multiple views to improve over conventional learning algorithms. Multiview learning has been studied as a potential framework to achieve such improved performance. Multiview methods operate with the assumption that different views cluster or label entities similarly. Such similarities have been exploited via *co-training* (Blum and Mitchell 1998) and *co-regularization* (Sindhwani and Rosenberg 2008). Co-training learns one hypothesis for each view which then bootstrap other views to converge to a coherent model (Blum and Mitchell 1998). Co-regularization, on the other hand, explicitly minimizes disagreement between views during training. Multiview methods have substantial theoretical and practical advantages over learning a single hypothesis by concatenating views (Nigam and Ghani 2000). For instance, Dasgupta et al. (2001) show that for semi-supervised multiview learning with two views, the probability of disagreement between views is an upper bound on the probability of error of either view's hypothesis.

Due to noisy measurements or unknown biases, different views may not cluster the entities similarly. To be robust to such mis-specification in model assumptions made by multiview clustering, we propose a method that maintains a separate posterior distribution for each view. In the proposed method, clustering coherence is imposed by encouraging posterior distributions of view-specific cluster memberships to be 'close' to each other, where closeness is measured via suitable divergences. Specifically, a weighted sum of divergences between current posterior estimates of cluster memberships is minimized. This co-regularization technique is combined with Expectation-Maximization (EM) (Dempster et al. 1977) to maximize the log-likelihood. The training process thus alternates between an inference phase that estimates and updates view-wise posterior distributions to encourage coherence followed by per-view parameter updates.

To specifically account for potential incoherence among views, we formulate the cost function as a weighted sum of Rényi divergences (Rényi 1960). Storkey et al. (2014) have observed that when aggregating opinions from biased experts or agents, the maximum entropy distribution is obtained via Rényi divergence aggregation (see Definition 1). An extreme case is when views don't agree on the cluster memberships, in which case, linear aggregation provides the best aggregate posterior. For instance, Bickel and Scheffer use linear aggregation in Co-EM (Bickel and Scheffer 2005), inadvertently assuming that different views mostly do not agree with respect to the cluster membership. Instead of assuming a bias free condition, we explore the utility of various aggregation strategies applied to the co-regularization framework. Hence, the proposed method can be applied with appropriate Rényi divergences best suited for different levels of discordance in view memberships. Co-EM (Bickel and Scheffer 2005) is a special case of our framework as it can be recovered as a specific setting of the Rényi divergence parameter for a fixed parametrization of weights as shown in Sect. 4.3. Extensive empirical evaluations are presented to demonstrate improved performance over existing multiview clustering methods as well as other methods of fusing information from multiple views.

Our main contributions are highlighted in the following:

- We propose a novel co-regularized multiview clustering algorithm that minimizes weighted sums of Rényi divergences.
- We show that an existing approach to probabilistic multiview clustering, namely Co-EM can be recovered as a special case of our framework.

– We present extensive empirical evaluation showing that the proposed class of methods significantly outperform strong baselines. Moreover, the choice of Rényi divergence can affect clustering performance, while simultaneously capturing biases in view-specific posterior cluster memberships. Empirical evaluation also demonstrates that our methods handle mixed data e.g. discrete and continuous data very well.

The rest of the manuscript is organized as follows. A brief survey of existing approaches to multiview clustering is provided in Sect. 2. Background and notation is given in Sect. 3. The proposed methods along with other modeling choices are detailed in Sect. 4. Extensive empirical evaluation on several data sets are in Sect. 5 followed by a discussion and conclusion in Sect. 6.

2 Related work

Related work in multiview unsupervised learning goes back to neural network models, a few of which are noted here. Becker and Hinton (1992), Schmidhuber and Prelinger (1993) maximize agreement between a given neural network module and a weighted output of its neighbors. De Sa and Ballard (1993) take advantage of complementary information available in different views by using separate modules for views feeding into a common output. Bickel and Scheffer (2005) introduced probabilistic multiview clustering using co-training.

Relatively recent models, like those proposed by Chaudhuri et al. (2009), Sa (2005) construct lower dimensional projections using multiple views of data. However, these methods are only applicable when at most two views are available. Kumar et al. (2010) and Tzortzis and Likas (2012) explore multiple kernel learning techniques where each view is represented as a kernel. Closely related to kernel techniques are multiview spectral clustering methods described in detail below.

Zhou and Burges (2007) propose a multiview spectral clustering method as a generalization of the normalized cuts algorithm. In a similar vein, Kumar and Daume-III (2011) update the similarity matrix of a given view based on the clustering of another view iteratively to produce a coherent clustering. Kumar et al. (2011) minimize disagreement between views by constraining the similarity matrices of views to be close in the Frobenius norm. While spectral methods are effective, they do not estimate cluster centroids, making interpretation and out-of-sample cluster assignments more challenging to implement. Our empirical studies show that proposed methods outperform spectral multiview clustering methods.

Further, connections between non-negative matrix factorization and clustering have also been utilized when multiple views are observed. For example, Liu et al. (2013) have shown that modeling user-feature matrices via multiview clustering based on non-negative matrix factorization (NMF) admits better empirical clustering performance compared to collective matrix factorization (Akata et al. 2011), a popular method for combining information from multiple sources. This further illustrates the advantage of multiview clustering over other related methods. Another approach for multi-view clustering using convex subspace representation learning has been proposed by Guo (2013). These methods estimate a subspace where different views are clustered similarly. Many of these methods, however, provide little insight into how views interact within the data. Probabilistic techniques such as ours are particularly useful when such exploration is required. Our empirical evaluation suggests improved performance of our models over NMF based multiview clustering. In addition, many models also deal with partially missing views (Eaton et al. 2014; Li et al. 2014) and demonstrate improved performance using multiview clustering. Lian et al. (2015) use a shared latent factor model to model heterogeneous multiview data and can also handle arbitrarily missing views, i.e., the case when a complete view may be missing for a sample. However, this model assumes a shared latent matrix across all views as opposed the proposed method which maintains separate cluster membership variables for each view. Our proposed methods can easily extend to handle missing views by simply not co-regularizing over the missing view.

Note that multiview clustering is distinct from cluster ensemble methods (Ghosh and Acharya 2011; Strehl and Ghosh 2003) that learn hypotheses for each view independently and find a consensus among the per-view results post-training. The latter methods do not share information during training and are thus more suitable for knowledge reuse (Strehl and Ghosh 2003).

One of the more popular applications of multiview clustering is to jointly model images and annotations, each constituting a view. The objective is to utilize annotations and images to learn the underlying clustering of images. This problem has been modeled in varied ways using unsupervised as well as supervised methods. We compare our multiview clustering framework to other relevant methods in the context of this application to motivate the differences in model assumptions. Recently, much supervised work has explored the utility of rich representations of label words and/or annotations in a high dimensional embedding space (Mikolov et al. 2013). A mapping is learned directly from the image view to the word embedding space (annotation view) so that relevant tags or labels are closer under some similarity metric (Frome et al. 2013; Akata et al. 2013) or ranked higher compared to the rest (Weston et al. 2011). Additionally, Akata et al. (2013) learn a mapping to a pre-defined attribute space to extend supervised image classification to unseen labels. Thus in this case, the target labeling is the same as the text or label view. In contrast, multiview clustering models aim to find the best underlying grouping of data jointly, thus differing in the underlying modeling assumptions. The multiview clustering methods presented in this paper are for completely unsupervised scenarios, and thus do not assume availability of labels for images. Further, the target clustering does not necessarily have to have a one-to-one mapping to the annotation views. Hence, in our empirical evaluation, we only compare our models to unsupervised multiview methods with similar modeling and data assumptions as ours.

3 Preliminaries

For non-negative integers K, vectors in $\mathbb{R}^K$ are denoted by lower-case bold (e.g., $\mathbf{x}$ with components $x_1, \ldots, x_K$). The set $\{1, 2, \ldots, K\}$ will be denoted $[K]$. The simplex Δ^K is the set:

$$\Delta^K = \left\{ \mathbf{a} \in \mathbb{R}^K \ : \ a_k \geq 0 \ \forall k \in [K], \ \sum_{k \in [K]} a_k = 1 \right\}$$

A *categorical distribution* is a discrete distribution over outcomes $\omega \in [K]$ parameterized by $\theta \in \Delta^K$ so that $Pr(\omega = k) = \theta_k$. It is a member of the exponential family of distributions. The natural parameters of categorical distribution are $\log \theta = (\log \theta_k)_{k \in [K]}$ and sufficient statistics are given by the vector of indicator functions for each outcome $\omega \in [K]$, denoted by $\mathbf{z}(\omega) \in \{0, 1\}^K$ with:

$$z_k(\omega) = \begin{cases} 1, & \text{if } \omega = k, \\ 0, & \text{otherwise.} \end{cases}$$

Given two categorical distributions $p(\omega)$ and $q(\omega)$, describing the distribution over the categorical random variable ω, the *divergence* of $p(\omega)$ from $q(\omega)$, denoted $\mathcal{D}(p(\omega) \| q(\omega))$, is a non-symmetric measure of the difference between the two probability distributions. The *Kullback-Leibler* or KL-divergence is a specific divergence denoted by $\mathrm{KL}(p(\omega)\|q(\omega))$ and is defined as follows.

KL-divergence of $p(\omega)$ from $q(\omega)$ is given by:

$$\mathrm{KL}(p(\omega)\|q(\omega)) = \mathrm{E}_{p(\omega)}\left[\log p(\omega) - \log q(\omega)\right] \tag{1}$$

This is also known as the relative entropy between $p(\omega)$ and $q(\omega)$. The relative entropy is non-negative and jointly convex with respect to both arguments. Further, we have that $\mathrm{KL}(p(\omega)\|q(\omega)) = 0$ iff $p(\omega) = q(\omega)$, for all ω.

The Rényi divergences (Rényi 1960) are a parametric family of divergences with many similar properties to the KL-divergence. Since our focus is on using these divergences to measure distances of distributions over cluster labels, we will focus on Rényi divergences for distributions over discrete random variables.

Definition 1 (van Erven and Harremoës 2012) Let p, q be two distributions for a random variable $\omega \in [K]$. The Rényi divergence of order $\gamma \in (0, 1) \cup (1, \infty)$ of $p(\omega)$ from $q(\omega)$ is,

$$\mathcal{D}_\gamma(p(\omega)\|q(\omega)) = \frac{1}{\gamma - 1}\log\left(\sum_{\omega=1}^{K} p(\omega)^\gamma q(\omega)^{(1-\gamma)}\right) \tag{2}$$

The definition may been extended for divergences of other orders like $\gamma = 0$, $\gamma \to 1$, and $\gamma \to \infty$ (van Erven and Harremoës 2012). Rényi divergences are non-negative $\forall \gamma \in [0, \infty]$. In addition, they are jointly convex in (p, q) $\forall \gamma \in [0, 1]$ and convex in the second argument q $\forall \gamma \in [0, \infty]$. As discussed in the comprehensive survey of Rényi divergences by van Erven and Harremoës (2012), many special cases of other commonly used divergences are recovered for specific choices of γ. For example, $\gamma = \frac{1}{2}$ and $\gamma = 2$ give Rényi divergences which are closely related to the Hellinger and χ^2 divergences, respectively, and the KL-divergence is recovered as a limiting case when $\gamma \to 1$. For the rest of the manuscript, we will abuse notation slightly and use $p(\omega)$ and $p(\mathbf{z})$ interchangeably to denote the same categorical distribution over outcomes in $[K]$.

4 Co-regularized multiview clustering using Rényi divergence minimization

We propose a co-regularization technique for multiview clustering using Rényi divergences. The generative model of the data is assumed to be a mixture model in each view. Let N be the total number of samples and V be the total number of views. For $n \in [N]$ and $v \in [V]$, let $\mathbf{x}_n^v$ represent the feature vector observed at view v for sample n. If the data sample n lies in cluster k in view v, the latent membership is indicated by the categorical random variable $\mathbf{z}_n^v \in \{0, 1\}^K$, where the k-th element of the vector, denoted by $z_{n,k}^v$ is 1 and the rest are 0. The vector $\boldsymbol{\pi}_n \in \Delta^K$ parametrizes the prior distribution over the categorical variable $\mathbf{z}_n^v$ and is the same for each view v. Each data sample is generated independent of the others. Also each view of a sample is generated independently conditioned on $\mathbf{z}_n^v$. Let the k^{th} cluster distribution at view v be parametrized by $\boldsymbol{\Psi}_k^v$. Let the set of all parameters for view v be denoted by $\boldsymbol{\Psi}^v$, i.e. $\boldsymbol{\Psi}^v = \{\boldsymbol{\Psi}_k^v\}$. The generative process can be represented by the plate model in Fig. 1 and is described as follows:

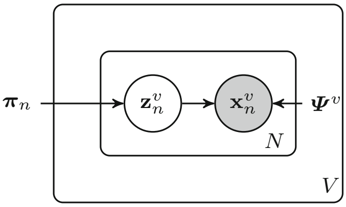

Fig. 1 Generative model for multi-view clustering

- For each n:
  - For each view v:
    - Choose $\mathbf{z}_n^v \sim p(\mathbf{z}_n^v; \boldsymbol{\pi}_n)$ categorical distribution parametrized by $\boldsymbol{\pi}_n$.
    - Choose $\mathbf{x}_n^v \sim p(\mathbf{x}_n^v | z_{n,k}^v = 1, \Psi_k^v)$ i.e., sample feature from the k^{th} cluster.

If no coherence conditions are imposed, each view can be modeled independently by maximizing the complete log-likelihood $\sum_{n \in [N]} \log p(\mathbf{x}_n^v, \mathbf{z}_n^v | \boldsymbol{\Psi}^v)$ using Expection Maximization (EM) (Dempster et al. 1977). Let $\boldsymbol{\Psi}_t^v$ be an estimate of the parameter $\boldsymbol{\Psi}^v$ at iteration t. The Expectation (E)-step, estimates posterior probabilities $p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\Psi}_t^v)\, \forall n \in [N], \forall v \in [V]$. A new estimate of cluster memberships can be obtained for each view, that decreases the Rényi divergence between all view-specific posteriors to encourage coherence. This core idea motivates the co-regularization technique proposed in the following subsections.

4.1 Global co-regularization

The proposed method minimizes a weighted sum of divergences between the current posterior or cluster membership estimates available at all views to estimate a new 'global' categorical distribution. We would like to trade-off between the 'global' posterior (accounting for co-regularization) and the view-specific unregularized posteriors. A new posterior distribution is estimated for every view $v \in [V]$ by minimizing the sum of divergences between the global categorical distribution and the view-specific posterior $p(\mathbf{z}_n^i | \mathbf{x}_n^i, \boldsymbol{\Psi}^i)\ \forall i \in [V]$.

At any iteration t, let $g_t(\mathbf{z}_n)$ be the global categorical posterior that is to be estimated from the independent posteriors $p(\mathbf{z}_n^i | \mathbf{x}_n^i, \vec{\Psi}_t^i)$. Let $\mathbf{w} \in \Delta^V$ denote a (known) weight vector of dimension V that determines the contribution of each view. Let the choice of divergence metric be known and given by γ. The i-th element of the vector $\mathbf{w}$ is denoted by w_i. The global distribution $g_t^*(\mathbf{z}_n)$ can be obtained by minimizing (3).

$$g_t^*(\mathbf{z}_n) = \underset{g(\mathbf{z}_n)}{\arg\min} \sum_{i \in [V]} \frac{w_i}{\gamma} \mathcal{D}_\gamma(p(\mathbf{z}_n^i | \mathbf{x}_n^i, \boldsymbol{\Psi}_t^i) || g(\mathbf{z}_n)) \tag{3}$$

To solve the weighted divergence minimization, a modified version of the variational algorithm proposed by Storkey et al. (2014) is used. Detailed derivation of the variational algorithm and specific updates are provided in Appendix 1.

For any given view v, it is desirable that its posterior be close to the global distribution $g_t^*(\mathbf{z})$ as well as its local estimate $p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\Psi}_t^v)$. Thus, we introduce a new weight vector $[w_g, (1 - w_g)] \in \Delta^2$, $0 < w_g < 1$ that achieves this trade-off. Let the co-regularized posterior obtained for the current view v being updated be denoted by $q_t(\mathbf{z}_n^v)$. Then, given γ and weights $[w_g, (1 - w_g)]$, $q_t(\mathbf{z}_n^v)$ can be estimated using:

$$q_t(\mathbf{z}_n^v) = \arg\min_{q(\mathbf{z}_n^v)} \frac{w_g}{\gamma} \mathcal{D}_\gamma(g^*{}_t(\mathbf{z}_n) || q(\mathbf{z}_n^v)) + \frac{(1 - w_g)}{\gamma} \mathcal{D}_\gamma(p(\mathbf{z}_n^v | \mathbf{x}_n^v, \boldsymbol{\Psi}_t^v) || q(\mathbf{z}_n^v)) \tag{4}$$

Algorithm 1 GRECO

```
Given data $\{\mathbf{x}_n^v\}$, $\gamma$, $\mathbf{w}$, Initialize $\boldsymbol{\pi}_n$, $\boldsymbol{\Psi}^v\ \forall v \in [V]$, $t = 0$
repeat
  for all $v \in [V]$ do
    //View specific E-step with the latest estimates of $\boldsymbol{\Psi}^i$s
    for all $i \in [V]$ do
      $p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}_t^i) \propto p(\mathbf{x}_n^i|\mathbf{z}_n^i, \boldsymbol{\Psi}_t^i)p(\mathbf{z}_n^i; \boldsymbol{\pi}_n)\ \forall n \in [N]$ in parallel
    end for
    //Coherence enforcing steps for current view $v$:
    Estimate $g_t^*(\mathbf{z}_n)$ by solving equation (3) using Algorithm 4 $\forall n \in [N]$ in parallel
    Estimate $q_t(\mathbf{z}_n^v)$ by solving equation (4) using Algorithm 5 $\forall n \in [N]$ in parallel
    //M-step for current view $v$:
    Using fixed responsibilities $q_t(\mathbf{z}_n^v)$,
    $\boldsymbol{\Psi}_{t+1}^v \leftarrow \arg\max_{\boldsymbol{\Psi}^v} \sum_{n\in[N]} \sum_{k\in[K]} q_t(\mathbf{z}_{n,k}^v) \log p(\mathbf{x}_n^v, z_{n,k}^v = 1; \boldsymbol{\Psi}^v)$
  end for
  $t \leftarrow t + 1$
until converged
```

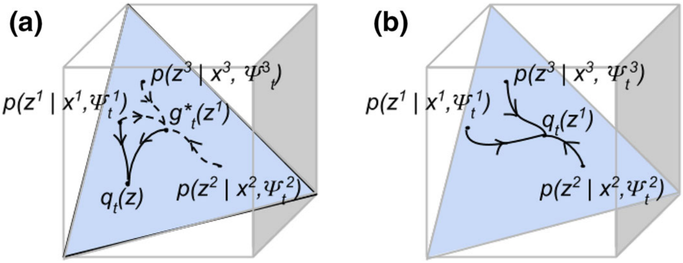
(a)
$p(z^1 \mid x^1, \Psi_t^1)$
$p(z^3 \mid x^3, \Psi_t^3)$
$g^*_t(z^1)$
$q_t(z)$
$p(z^2 \mid x^2, \Psi_t^2)$
(b)
$p(z^1 \mid x^1, \Psi_t^1)$
$p(z^3 \mid x^3, \Psi_t^3)$
$q_t(z^1)$
$p(z^2 \mid x^2, \Psi_t^2)$

Fig. 2 **a** *Left* Inner loop of inference in GRECO for $V = 3$ and $K = 3$ at iteration t. The inference shows updates for view 1. $p(z^i|x^i, \boldsymbol{\Psi}_t^i)$ for $i \in [3]$ are the independent view-wise posteriors, $g_t^*(z)$ is the global distribution for GRECO. $q_t(\mathbf{z}^1)$ is the re-estimated posterior for view 1. **b** *Right* Inner loop of inference for LYRIC

Given $g_t^*(\mathbf{z}_n)$, (4) can be solved to minimize the weighted sum of two divergences only. Since all views admit separate aggregate posteriors, the M-step can be executed for each view independent of all other views using the conventional M-step for a mixture model. In order that all the views are in coherence with the latest posterior beliefs of all other views, an M-step for the view should follow every view's posterior update. This prevents the algorithm from getting stuck at local minima by avoiding parameter estimates that may agree with old and potentially discordant beliefs of other views. For specific updates in the M-step based on the chosen mixture model, please refer to Appendix 4.

The above procedure of inferring per-view responsibilities independently, followed by a coherence enforcing step using (3) and (4) and a view-specific M-step is computed for each view. The complete algorithm for the proposed global co-regularization, called GRECO (Global REnyi divergence based CO-regularization) and is provided by Algorithm 1. The inference step for GRECO, specifically the inner loop update is shown in Fig. 2a for a toy problem with three views for $K = 3$. Note that (3) and (4) can be computed in parallel for data samples $n \in [N]$. This is because our target co-regularized posterior is independent for each sample (and each view) and can be factored in the product form over samples as well as views. A detailed proof of how that leads to an embarrassingly parallel co-regularization algorithm is provided in Appendix 2.

4.2 Local co-regularization

We now consider two limiting cases, when $w_g = 0$ and $w_g = 1$ in (5). The first case ($w_g = 0$) is trivial as it does not use co-regularization at all and is therefore equivalent to the ensemble method (Strehl and Ghosh 2003). The latter recovers a new method. We consider the non-trivial case when $w_g = 1$ separately for several reasons. First, we are able recover an existing multiview clustering algorithm, Co-EM as a special case of this setting for a certain choice of Rényi divergences ($\gamma \to 1$). Thus Co-EM is also a special case of our most general setting, GRECO for $\gamma \to 1$ and $w_g = 1$. Further, our empirical evaluation suggests better performance of the most general case of the proposed method (GRECO) in most cases as opposed to this special case ($w_g = 1$) suggesting that a non-trivial trade-off between the global posterior and the view-specific unregularized posterior in the E-step is advantageous. A useful analogy we would like to draw here is between Gaussian mixture models versus the k-means algorithm, used for soft and hard clustering respectively. The latter is a limiting case of the former (as widths go to zero, and using identical, isotropic covariances) but is considered as a separate algorithm because of its special properties. The co-regularization framework in GRECO with $w_g = 1$ does not involve an additional trade-off between the global posterior and the unregularized view-specific posterior. The minimizer of (5) in this case is exactly equal to $g^*(\mathbf{z}_n)$. Thus, the view-specific co-regularized posterior $q(\mathbf{z}_n^v)$ is equal to the global posterior $g^*(\mathbf{z}_n)$. Note that in a given iteration, only view v is co-regularized so that $q(\mathbf{z}_n^v) = g^*(\mathbf{z}_n)$. All other views are not updated in the inference and the learning step in the same iteration. The procedure is repeated subsequently for all views. We call this algorithm LYRIC (LocallY weighted Rényi dIvergence Co-regularization). As in GRECO, the outer loop of LYRIC iterates over each view v and the inner loop carries out a coherence enforcing E-step for the given view followed by an M-step. The E-step comprises of estimating independent view-specific posteriors followed by a *local* co-regularization step that updates the current view's posterior. It is important to highlight that LYRIC does not result in the same estimates as GRECO every iteration. This is because view-specific posteriors will be different in each iteration for GRECO and LYRIC owing to the different stages of co-regularization. The details of local co-regularization (LYRIC) are now explained in the following.

Let v be the current view to be updated at any iteration t and let $q_t(\mathbf{z}_n^v)$ denote the newly obtained posterior at view v for sample n. Local co-regularization solves the coherence equation given by (5).

$$q_t(\mathbf{z}_n^v) = \underset{q(\mathbf{z}_n^v)}{\arg\min} \sum_{i \in [V]} \frac{w_i}{\gamma} \mathcal{D}_\gamma(p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}_t^i)||q(\mathbf{z}_n^v)) \tag{5}$$

Similar to GRECO, a per-view M-step can now be executed to update per-view parameters according to the modified responsibilities. The procedure is repeated iteratively for all views $v \in [V]$. The final algorithm, LYRIC, is illustrated in Algorithm 2. Figure 2b shows the inference step for a single view in LYRIC.

4.3 Special case I: $\gamma \to 1$

If γ is chosen such that $\gamma \to 1$, the minimizer of weighted sum of Rényi divergences admits a closed form solution. Specifically, $\gamma \to 1$ reduces the cost to a weighted sum of KL-divergences with the target distribution on the right hand side of KL-divergence (Storkey et al. 2014). Consider (5), for instance, with $\gamma \to 1$. Let the per-view posterior, $p(\mathbf{z}^i|\mathbf{x}^i, \boldsymbol{\Psi}^i)$

Algorithm 2 LYRIC

```
Given data $\{\mathbf{x}_n^v\}$, $\gamma$, $\mathbf{w}$, Initialize $\boldsymbol{\pi}_n$, $\boldsymbol{\Psi}^v\ \forall v \in [V],\ t = 0$
repeat
  for all $v \in [V]$ do
    //View specific E-steps with the latest parameter estimates of $\boldsymbol{\Psi}^i$s
    for all $i \in [V]$ do
      $p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}_t^i) \propto p(\mathbf{x}_n^i|\mathbf{z}_n^i, \boldsymbol{\Psi}_t^i)p(\mathbf{z}_n^i; \boldsymbol{\pi}_n)\ \forall n \in [N]$ in parallel
    end for
    //Coherence enforcing step for current view $v$:
    Estimate $q_t(\mathbf{z}_n^v)$ with Eq. 5 using Algorithm (5) $\forall n \in [N]$ in parallel
    //M-step for current view $v$:
    Using fixed responsibilities $q_t(\mathbf{z}_n^v)$,
    $\boldsymbol{\Psi}_{t+1}^v \leftarrow \arg\max_{\boldsymbol{\Psi}^v} \sum_{n \in [N]} \sum_{k \in [K]} q_t(\mathbf{z}_{n,k}^v) \log p(\mathbf{x}_n^v, z_{n,k}^v = 1; \boldsymbol{\Psi}^v)$
  end for
  $t \leftarrow t + 1$
until converged
```

be parametrized by $\boldsymbol{\theta}^i \in \Delta^K$. Let the target distribution $q(\mathbf{z}_n^v)$, be parametrized by $\boldsymbol{\phi}^v \in \Delta^K$. The cost function given by (5) can be simplified to (6).

$$q(\mathbf{z}^v) = \underset{q(\mathbf{z}^v)}{\arg\min} \sum_{i \in [V]} w_i \mathrm{KL}\Big(p(\mathbf{z}^i|\mathbf{x}^i, \boldsymbol{\Psi}^i) \| q(\mathbf{z}^v)\Big) \tag{6}$$

For categorical distributions, the closed form solution of (6) is given by (7) as was derived by Garg et al. (2004). Refer to Appendix 3 for a proof.

$$\boldsymbol{\phi}^v = \sum_{i \in [V]} w_i \boldsymbol{\theta}^i \tag{7}$$

Note that the linear aggregation closed form is not specific to LYRIC and can be generalized to GRECO for the choice of $\gamma \to 1$. In GRECO, both (3) and (4) reduce to linearly aggregating over per-view posteriors in the former and weighted divergence minimization between the global posterior and the current view's posterior in the latter case.

Specifically, if $w_v = (1 - \alpha)$ for the view v currently being updated, and $w_i = \frac{\alpha}{V-1}$, where $0 \leq \alpha \leq 1$ for $i \neq v, i \in [V]$, the LYRIC algorithm recovers Co-EM when $\gamma \to 1$. Thus Co-EM is a special case of LYRIC.

4.4 Special case II: $\gamma \to 0$

When $\gamma \to 0$, (5) has been shown by Storkey et al. (2014) to be equivalent to a minimization over a weighted sum of the KL-divergences with the target distribution as the argument on the left-hand side of KL-terms. The closed form solution in this case is an averaging of the parameters $\boldsymbol{\theta}^i\ \forall i \in [V]$ in the *log*-space weighted by $w_i\ \forall i \in [V]$ (Garg et al. 2004) as shown in (8). The proof is detailed in Appendix 3.

$$\log \boldsymbol{\phi}^v = \sum_{i \in [V]} w_i \log \boldsymbol{\theta}^i \tag{8}$$

This result is also general and applicable to (3) and (4) with appropriate weighting. For these special cases, the variational updates can be avoided to use the simpler closed form updates for GRECO and LYRIC. Note that (8) can be equivalently written as:

$$\boldsymbol{\phi}^v = \prod_{i \in [V]} {\boldsymbol{\theta}^i}^{w_i} \tag{9}$$

This further suggests that when $\gamma \rightarrow 1$, the parameters across views contribute equally owing to linear averaging as opposed to when $\gamma \rightarrow 0$ (9) where extreme values of the posteriors may dominate. Conventionally, a product of experts model (Hinton 2002; Storkey et al. 2014) uses such a product to combine beliefs from independently trained models, for example in an ensemble setting.

Note that co-regularization in each GRECO and LYRIC adds an additional complexity of $\mathcal{O}(NKV^2)$ per iteration where N is the sample size, K is the number of clusters and V is the number of views, compared to the unregularized method. As suggested before, the operations can be trivially parallelized over data samples as well as for calculations required to estimate unnormalized variational parameters for each cluster (see Appendices 1, 2). For the case where all views are Gaussian mixtures, the complexity per outer iteration is $\mathcal{O}(NKV^2T_{inner} + NKV + \sum_{v\in[V]} d_v^2 K)$ where T_{inner} is the number of inner iterations for variational estimation of co-regularized posteriors, d_v is the dimension of view v. In each of the special cases described earlier, i.e. when $\gamma \rightarrow 0$ and $\gamma \rightarrow 1$, the complexity reduces to $\mathcal{O}(NKV + \sum_{v\in[V]} d_v^2 K)$ per iteration, same as that of Co-EM, due to closed form solutions available for co-regularization. In the general case, the largest source of computational overhead in the proposed algorithm is due to the variational procedure currently employed to impose co-regularization. However, we are not bound to such a procedure and any accelerated methods available for solving (3), (4) and (5) can be adopted, if available. Further, our variational procedure is trivially parallelizable over samples (see Appendix 2 for relevant proof) whereas co-regularization/co-training techniques for the baselines (see Sect. 5.1) are not. This allows us to improve training efficiency to scale to large datasets.

4.5 Choice of weights and Rényi divergences

For empirical studies, we parametrize the weights for easy comparison with baselines. Let $0 \le \alpha \le 1$ be a scalar. For every view $v \in [V]$ being updated, $w_v = 1-\alpha$. For all other views, $w_i = \frac{\alpha}{V-1}\ \forall i \in [V],\ i \neq v$. At every stage in the outer loop of either GRECO or LYRIC, the current view being updated is weighted by $1-\alpha$ and the rest are weighted equally $\frac{\alpha}{V-1}$. This also ensures fair comparison with Co-EM by maintaining the same parametrization of weights. Therefore, all experiments demonstrate that a significant boost in clustering performance can be obtained via a suitable choice of Rényi divergences. We evaluated the performance of GRECO and LYRIC for different choices of α and γ. Section 5.3 shows the performance of the model across different choices of the divergence parameter. Specifically, for comparison with baselines, we choose the best performing set of α and γ based on average accuracy of hold-out clustering assignment across five trials.

4.6 Prediction on hold-out samples

For out-of-sample cluster assignment, the conventional E-step with the learned parameters is used to obtain per-view posteriors for a test sample for all views independently. It is now desirable to obtain one aggregate posterior $q(\mathbf{z})$ as follows.

$$q(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \sum_{v\in[V]} w_v \mathcal{D}_\gamma(p(\mathbf{z}^v|\mathbf{x}^v, \boldsymbol{\Psi}^{v*})||q(\mathbf{z})) \tag{10}$$

For LYRIC, a global posterior can then be obtained using (10) for a given choice of γ and $\mathbf{w}$ (see Sect. 4.5) and the set of corresponding learned parameters from LYRIC $\boldsymbol{\Psi}^{v*}$. Similarly for GRECO, the E-step is run for all views independently followed by executing (10) to obtain a global posterior. A hard clustering is simply the MAP assignment of $\mathbf{z}$ w.r.t. the distribution

$q(\mathbf{z})$. Empirical performance of LYRIC at $\gamma \to 1$ can differ from Co-EM due to different methods of obtaining the consensus clustering. Specifically, let $\boldsymbol{\pi}_n^c$ and $\boldsymbol{\Psi}^{vc}, \forall v \in [V]$ be the estimates of the prior distribution parameter and mixture model parameters learned by Co-EM respectively. Then the consensus clustering distribution and the corresponding MAP assignment w.r.t. the consensus distribution in Co-EM is given by (11) for a data sample $\mathbf{x} \equiv \{\mathbf{x}^v, \forall v \in [V]\}$,

$$\arg\max_{k\in[K]} \frac{\pi_k^c \prod_{v\in[V]} p(\mathbf{x}^v|\boldsymbol{\Psi}^{vc}, \mathbf{z}_k = 1)}{\sum_{k'\in[K]} \pi_{k'}^c \prod_{v\in[V]} p(\mathbf{x}^v|\boldsymbol{\Psi}^{vc}, \mathbf{z}_{k'} = 1)} \tag{11}$$

Note that this method of obtaining a consensus clustering used by Co-EM is equivalent to the E-step of a multiview latent variable model that shares a single latent clustering variable across all views. As opposed to Co-EM, GRECO and LYRIC obtain a consensus via linear aggregation for $\gamma \to 1$ and weights $\mathbf{w}$ as shown in (12).

$$\arg\max_{k\in[K]} \sum_{v\in[V]} w_v p(z_k^v = 1|\mathbf{x}^v, \boldsymbol{\Psi}^{v*}) \tag{12}$$

5 Experiments

The proposed methods have been extensively compared with existing multiview clustering models to show that the choice of divergence obtained by tuning γ is of significance, as well as to demonstrate that Rényi divergence is a reasonable choice for co-regularization. All datasets were trained using both LYRIC and GRECO algorithms for different values of $\gamma \in [0, 1]$ discretized in the corresponding log-space. Very high values of Rényi divergences did not matter significantly affect the performance. The weights $\mathbf{w}$ are reparametrized as described in Sect. 4.5. For all datasets, ground-truth cluster labels are known and utilized for objective evaluation and comparison to baselines. All models and baselines were trained on the same training and hold-out data for five trials with best performing models chosen based on average clustering accuracy for comparison purposes. The mapping between cluster labels to ground truth labels is solved using Hungarian matching (Kuhn 1955). For comparison to baselines, we only report the best performance obtained across different choices of $\mathbf{w}$ and γ. Hold-out assignment results have only been compared to baselines that explicitly mention a mechanism to obtain hold-out cluster assignment and empirically test the same. We report Clustering Accuracy, Precision, Recall, F-measure, NMI (Strehl and Ghosh 2003) and Entropy (Bickel and Scheffer 2005) for our evaluation. Lower entropy is better while higher values of other metrics show a better performing algorithm. All metrics are defined in Appendix 5. Note that the empirical evaluation here maintains prior cluster distribution $\boldsymbol{\pi}_n$ to be equal for all samples n for all probabilistic models, including GRECO and LYRIC without loss of generality. Results demonstrating empirical convergence for a sample fold with multiple initializations (in negative log-likelihood) of GRECO and LYRIC have been included in Appendix 6.[1] To the best of our knowledge, our empirical evaluation is the most extensive evaluation of multiview clustering methods compared to prior work in terms of the number of datasets, number of views and comparison to existing baselines.

[1] For the CUB dataset, we only have results with a single initialization for a single train-test split. However, average over different splits shows the same trend.

5.1 Baselines

The proposed methods are compared to an extensive set of baselines. The baselines are briefly described here.

- *Shared Latent Variable Model (Joint)*: An alternative way of modeling multiple views is to have one latent variable that denotes the cluster membership across all views. This is called the 'Joint' model. This model is equivalent to concatenating views especially in the most commonly assumed scenario i.e. all views are Gaussian mixtures with diagonal covariances.
- *Ensemble Clustering Model (Ensemble)* (Strehl and Ghosh 2003): This model trains each view independently followed by a consensus evaluation. To predict the hard clustering assignment, the label correspondence among views is obtained using Hungarian matching (Kuhn 1955). A single posterior is obtained using the same equation as (10) with KL-divergence (log-aggregation), followed by a MAP assignment. This method is compared to only when at most two views are available.
- *Co-EM* (Bickel and Scheffer 2005): Co-EM estimates a mixture model per view subject to cross-entropy constraints. The weights for each view are parametrized by $\eta \in [0, 1]$ and the results corresponding to the best performing η are reported.
- *Co-regularized Spectral Clustering (Co-reg (Sp))* (Kumar et al. 2011): This is the state-of-the-art spectral multiview clustering. The results corresponding to the best performing λ parameter (between 0.01 to 0.1 as suggested by authors) are reported. The implementation provided by the authors is used.[2]
- *Minimizing Disagreement (Min-dis (Sp))* (Sa 2005): This is another spectral clustering technique proposed by (Sa 2005) for 2 views only. The implementation used was implemented and compared to by Kumar et al. (2011).
- *CCA for Mixture Models (CCA-mvc)* (Chaudhuri et al. 2009): This method uses Canonical Correlation Analysis to project views on a lower dimensional space. This model can be used for 2 views only.
- *NMF based Multiview Clustering (NMF-mvc)* (Liu et al. 2013): This method uses non-negative matrix factorization for multiview clustering. The original implementation provided by the authors was used for empirical evaluation.[3]

A k-means clustering algorithm is used independently for each view to initialize distribution parameters for all probabilistic models. An approximate Hungarian matching problem is solved using the k-means cluster assignments for initialization.

5.2 Datasets

The datasets are chosen referencing prior work in multiview clustering. Details of the datasets are provided in the following.

- *Twitter multiview*[4] (Greene and Cunningham 2013): This is a collection of twitter datasets in five topical areas (politics-UK, politics-Ireland, Football etc.). Each user has views corresponding to users they follow, their followers, mentions, tweet content etc. We use the politics-uk dataset with three views (mentions, re-tweets and follows). The labels correspond to one of five party memberships of each user. Each view is a *bag-of-words* vector and modeled as a mixture of multinomials for the probabilistic models.

[2] http://www.umiacs.umd.edu/~abhishek/papers.html.

[3] http://jialu.cs.illinois.edu/.

[4] http://mlg.ucd.ie/aggregation/.

- *WebKB*[5]: This dataset consists of webpage information from four university websites: Cornell, Texas, Washington and Wisconsin. We show results for the Cornell dataset. Each sample is a webpage with two views, one view of which is the text content (*bag-of-words*) format and web-links into and out of the webpage (binary *bag-of-words* vector). Each webpage can be clustered into one of five topics. Each view is modeled as a mixture of multinomials.
- *NUS-Wide Object*[6] (Chua et al. 2009): This dataset consists of 31 object classes. Of these, we sub-sample in a balanced manner for 10 classes, with 50 samples belonging to each class. We use 6 views, namely edge histograms (mixture of Gaussians), *bag-of-visual words* of SIFT features (mixture of multinomial distributions) and normalized correlogram (mixture of Gaussians), color histogram (mixture of multinomials), wavelet texture (mixture of Gaussians) and block-wise color moments (mixture of Gaussians).
- *CUB-200-2011*[7] (Wah et al. 2011): This dataset consists of 200 classes and 11,800 data samples. We use the binary attributes and Fisher Vector representations of images as our views. The binary attributes are modeled as mixture-of-multinomials and the Fisher vectors as Gaussian mixtures. We assume diagonal covariances for all views modeled as a mixture of Gaussians in all datasets.

5.3 Results

Tables 1, 2, 3 and 4 show clustering and out-of-sample cluster assignment results for the datasets mentioned in Sect. 5.2 in that order. Note that results are marked NA if any of the baseline methods were not extendable to more than two views or could not be compared due to limiting model assumptions e.g. non-negativity required by NMF-mvc (Liu et al. 2013). The tables only consist of results corresponding to the γ parameter that provided the best results across different choices of γ for both GRECO and LYRIC on a hold-out dataset. Additionally, Fig. 3 shows performance of GRECO and LYRIC using different Rényi divergences parametrized by $\log(\gamma)$ in comparison with Co-EM, that uses linear aggregation, corresponding to $\gamma \to 1$. The performance across different γ provides further insights into performance of the proposed co-regularization method.

The proposed methods outperform almost all the baselines consistently across different datasets. In addition, hold-out cluster assignment performance is better for both models across most datasets. Improved performance over ensemble methods suggests co-regularization improves on the view-wise clustering approaches. In addition, results also suggest that sharing a single latent variable (see Joint Model) across views is restrictive. In the low bias regime, GRECO has particular advantages over LYRIC because of the additional trade-off in regularization. When the bias across views is low, the additional regularization potentially accelerates convergence by restricting the deviation from view-specific unregularized posteriors, especially when initial model parameters may be noisy. In the high bias case, LYRIC shows some advantage (see Table 2-WebKB data). It is important to note that overall, the general trend of performance of both GRECO and LYRIC is consistent for each dataset (see Fig. 3). In particular, the performance peaks for the most appropriate choice of γ that best captures inherent biases across views for both algorithms for all datasets and this choice of divergence is the same for GRECO as well as LYRIC.

[5] http://lig-membres.imag.fr/grimal/data.html.

[6] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

[7] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

Table 1 Twitter data (politics-uk, 3 views), best results obtained for $\gamma = 0.01$ for GRECO and LYRIC

Clustering results							
Method	Accuracy	Precision	Recall	F-measure	NMI	Entropy	Time (s)
GRECO	**0.9075 (0.0201)**	**0.9403 (0.0316)**	**0.8713 (0.0366)**	**0.9039 (0.0217)**	**0.7887 (0.0478)**	**0.2971 (0.1001)**	7.1241 (1.0122)
LYRIC	**0.886 (0.0284)**	**0.9601 (0.01)**	0.8441 (0.0596)	**0.8976 (0.0372)**	**0.8045 (0.0403)**	**0.2434 (0.0431)**	7.0888 (0.9755)
Co-EM	0.8346 (0.0488)	0.8973 (0.0346)	0.7559 (0.0757)	0.8197 (0.0566)	0.7058 (0.0406)	0.3876 (0.0536)	2.3714 (0.9746)
Joint	0.7893 (0.0491)	0.7737 (0.0792)	0.7167 (0.0679)	0.7413 (0.0535)	0.5806 (0.053)	0.6497 (0.11)	**0.3623 (0.1047)**
Ensemble	NA	NA	NA	NA	NA	NA	NA
Co-reg (Sp)	0.557 (0.0221)	0.7122 (0.0215)	0.4326 (0.0197)	0.5382 (0.0213)	0.5079 (0.018)	0.6329 (0.0293)	**1.6324 (0.1538)**
CCA-mvc	NA	NA	NA	NA	NA	NA	NA
Min-dis (Sp)	NA	NA	NA	NA	NA	NA	NA
NMF-mvc	0.4418 (0)	0.3802 (0)	**0.972 (0)**	0.5466 (0)	0.0161 (0)	1.5769 (0)	6.0709 (0.0895)
Hold-out cluster assignment results							
GRECO	**0.9238 (0.0136)**	**0.9047 (0.0384)**	**0.9021 (0.0559)**	**0.9022 (0.0307)**	**0.7784 (0.0418)**	**0.3417 (0.0703)**	NA
LYRIC	**0.8452 (0.0854)**	**0.8537 (0.0283)**	**0.8123 (0.1353)**	**0.8291 (0.0888)**	**0.6803 (0.0635)**	**0.4438 (0.0697)**	NA
Co-EM	0.781 (0.0287)	0.8282 (0.0406)	0.6735 (0.0425)	0.7425 (0.0379)	0.5916 (0.0566)	0.4988 (0.1454)	NA
Joint	0.769 (0.0644)	0.6629 (0.064)	0.7093 (0.1399)	0.6797 (0.0828)	0.4916 (0.0829)	0.7895 (0.1495)	NA
Ensemble	NA	NA	NA	NA	NA	NA	NA
CCA-mvc	NA	NA	NA	NA	NA	NA	NA

Ensemble model, CCA-mvc, Min-dis (Sp) can cluster at most two views and marked NA otherwise. Co-reg (Sp), Min-dis (Sp) and NMF-mvc do not explicitly compare hold-out cluster assignment results and have not been compared to for hold-out assignment performance. Top two methods w.r.t. each metric are highlighted

Table 2 Cornell (WebKB 2-views), best results obtained for $\gamma = 0.1$ for GRECO and $\gamma \to 1$ for LYRIC

Clustering results							
Method	Accuracy	Precision	Recall	F-measure	NMI	Entropy	Time (s)
GRECO	**0.5859 (0.0148)**	**0.431 (0.0385)**	0.617 (0.0451)	**0.5066 (0.0327)**	0.2747 (0.0145)	1.5578 (0.0379)	0.3404 (0.0323)
LYRIC	**0.5885 (0.0254)**	0.4135 (0.0351)	**0.6591 (0.0292)**	**0.5075 (0.0295)**	**0.2771 (0.024)**	1.5697 (0.0489)	0.3174 (0.03)
Co-EM	0.5269 (0.0325)	0.3753 (0.0324)	0.5485 (0.0777)	0.4432 (0.0374)	0.1908 (0.0187)	1.7216 (0.0366)	0.8036 (0.1299)
Joint	0.4179 (0.025)	0.3232 (0.0184)	0.4805 (0.0849)	0.3846 (0.0334)	0.1405 (0.0084)	1.8257 (0.0051)	0.1855 (0.0178)
Ensemble	0.5064 (0.0304)	0.3535 (0.0199)	0.6008 (0.1335)	0.4376 (0.0326)	0.2099 (0.0352)	1.7026 (0.0592)	**0.0341 (0.0011)**
Co-reg (Sp)	0.5551 (0.0494)	**0.5083 (0.0157)**	0.4596 (0.0354)	0.4824 (0.0252)	**0.3929 (0.0167)**	**1.2719 (0.0386)**	2.065 (0.0201)
CCA-mvc	0.4526 (0.014)	0.3118 (7e−04)	0.4751 (0.0304)	0.3762 (0.0095)	0.1665 (0.0019)	**1.2664 (0.0903)**	**0.0786 (0.0188)**
Min-dis (Sp)	0.3756 (0.0154)	0.32 (0.0023)	0.3116 (0.0524)	0.3139 (0.0251)	0.1614 (0.0048)	1.7744 (0.0207)	0.0904 (0.0366)
NMF-mvc	0.4103 (0)	0.2606 (0)	**0.9605 (0)**	0.41 (0)	0.0569 (0)	2.0497 (0)	5.6911 (0)
Hold-out cluster assignment results							
GRECO	0.4513 (0.0739)	0.2995 (0.051)	0.5782 (0.1985)	0.3872 (0.0795)	0.1777 (0.0573)	1.7211 (0.147)	NA
LYRIC	**0.5026 (0.0693)**	**0.3493 (0.0683)**	0.5541 (0.2054)	**0.4238 (0.1034)**	**0.2223 (0.1096)**	**1.63 (0.2153)**	NA
Co-EM	0.4205 (0.0862)	0.2788 (0.0606)	0.538 (0.1851)	0.3626 (0.0908)	0.1762 (0.035)	1.7269 (0.0966)	NA
Joint	0.4564 (0.0585)	0.2861 (0.0344)	**0.6214 (0.0806)**	0.39 (0.0391)	0.1934 (0.0583)	1.7096 (0.0844)	NA
Ensemble	**0.5487 (0.1082)**	**0.4123 (0.1742)**	**0.7356 (0.107)**	**0.5016 (0.1051)**	**0.2981 (0.1633)**	**1.5027 (0.407)**	NA
CCA-mvc	0.4103 (0)	0.3103 (0.007)	0.4 (0.0123)	0.3494 (0.0074)	0.1192 (0.0191)	1.7107 (0.0361)	NA

Table 3 NUSWideObj Dataset (6 views), best results obtained for $\gamma = 0.1$ for GRECO and LYRIC

Clustering results							
Method	Accuracy	Precision	Recall	F-measure	NMI	Entropy	Time (s)
GRECO	**0.3805 (0.0089)**	**0.245 (0.0058)**	**0.3362 (0.0347)**	**0.2829 (0.0146)**	**0.3276 (0.0199)**	**2.2687 (0.0574)**	8.0385 (1.2579)
LYRIC	**0.3805 (0.0089)**	**0.245 (0.0058)**	**0.3362 (0.0347)**	**0.2829 (0.0146)**	**0.3276 (0.0199)**	**2.2687 (0.0574)**	8.0099 (1.2586)
Co-EM	0.347 (0.0118)	0.2171 (0.011)	0.3006 (0.0184)	0.2518 (0.0092)	0.2903 (0.0089)	2.3918 (0.0319)	4.3041 (0.7188)
Joint	0.3115 (0.0151)	0.1882 (0.016)	**0.346 (0.0303)**	0.2437 (0.0202)	0.2454 (0.0157)	2.5884 (0.0481)	**2.8231 (0.7605)**
Ensemble	NA	NA	NA	NA	NA	NA	NA
Co-reg (Sp)	0.3785 (0.0202)	0.2629 (0.0128)	0.2816 (0.0196)	0.2718 (0.0153)	0.318 (0.0162)	2.273 (0.0531)	**2.5275 (0.0541)**
CCA-mvc	NA	NA	NA	NA	NA	NA	NA
Min-dis (Sp)	NA	NA	NA	NA	NA	NA	NA
NMF-mvc	NA	NA	NA	NA	NA	NA	NA
Hold-out cluster assignment results							
GRECO	**0.412 (0.0409)**	**0.225 (0.0228)**	**0.3369 (0.0177)**	**0.2691 (0.017)**	**0.4178 (0.0246)**	**1.9934 (0.0893)**	NA
LYRIC	**0.412 (0.0409)**	**0.225 (0.0228)**	**0.3369 (0.0177)**	**0.2691 (0.017)**	**0.4178 (0.0246)**	**1.9934 (0.0893)**	NA
Co-EM	0.372 (0.0217)	0.2074 (0.0232)	0.2964 (0.0405)	0.2437 (0.0289)	0.3975 (0.026)	2.052 (0.0856)	NA
Joint	0.334 (0.0241)	0.1806 (0.019)	0.352 (0.0374)	0.2387 (0.0248)	0.329 (0.0294)	2.3533 (0.092)	NA
Ensemble	NA	NA	NA	NA	NA	NA	NA
CCA-mvc	NA	NA	NA	NA	NA	NA	NA

Since this data has three views that take negative values, we do not compare against NMF-mvc. CCA-mvc and Min-dis (Sp) cannot be extended for more than two views

Table 4 CUB-200-2011 (2 views), best results obtained for $\gamma \rightarrow 1$ for GRECO and LYRIC

Clustering results							
Method	Accuracy	Precision	Recall	F-measure	NMI	Entropy	Time (s)
GRECO	**0.2231 (0.0039)**	**0.1052 (0.0034)**	**0.1757 (0.005)**	**0.1316 (0.0038)**	**0.5109 (0.006)**	**3.8498 (0.0508)**	2255.5 (169.34)
LYRIC	0.2189 (0.0061)	0.099 (0.004)	**0.1748 (0.005)**	**0.1264 (0.0036)**	**0.5071 (0.0051)**	3.8867 (0.045)	2069.6 (143.74)
Co-EM	0.0939 (0.0104)	0.0111 (0.0014)	0.0891 (0.0135)	0.0197 (0.0019)	0.301 (0.0146)	5.5905 (0.1318)	3355.9 (2382.4)
Joint	0.0715 (0.0035)	0.0109 (2e−04)	0.0582 (0.0035)	0.0183 (2e−04)	0.2473 (0.0063)	5.9822 (0.0511)	2004.1 (124.85)
Ensemble	0.0432 (9e−04)	0.0084 (3e−04)	0.0809 (0.0119)	0.0151 (3e−04)	0.1756 (0.0067)	6.5442 (0.0589)	767.78 (56.32)
Co-reg (Sp)	0.2118 (0.0081)	**0.1031 (0.0042)**	0.118 (0.0053)	0.11 (0.0046)	0.4896 (0.0059)	3.9224 (0.0431)	901.21 (11.716)
CCA-mvc	**0.2213 (0.007)**	0.0759 (0.0066)	0.1527 (0.0069)	0.1012 (0.006)	0.5003 (0.0038)	**3.4551 (0.0454)**	**4.8814 (0.1651)**
Min-dis (Sp)	0.1994 (0.0093)	0.0795 (0.0043)	0.1214 (0.0077)	0.0961 (0.0054)	0.4691 (0.0055)	4.1377 (0.0408)	**594.78 (20.514)**
NMF-mvc	NA	NA	NA	NA	NA	NA	NA
Hold-out cluster assignment results							
GRECO	**0.2133 (0.0078)**	**0.0601 (0.0046)**	**0.1304 (0.008)**	**0.0822 (0.0057)**	**0.57 (0.0048)**	**3.4714 (0.0417)**	NA
LYRIC	0.2066 (0.0085)	0.0531 (0.0027)	0.1284 (0.0045)	0.0751 (0.0025)	0.5644 (0.0043)	3.5276 (0.0417)	NA
Co-EM	0.0712 (0.0712)	0.0086 (0.0086)	0.1208 (0.1208)	0.0159 (0.0159)	0.3347 (0.02)	5.5129 (0.1822)	NA
Joint	0.0603 (0.0603)	0.0093 (0.0093)	0.0671 (0.0671)	0.0163 (0.0163)	0.3259 (0.0116)	5.5296 (0.0935)	NA
Ensemble	0.0508 (0.0508)	0.0088 (0.0088)	0.09 (0.09)	0.016 (0.016)	0.2808 (0.0182)	5.8884 (0.1399)	NA
CCA-mvc	**0.2512 (0.0064)**	**0.0727 (0.0048)**	**0.1444 (0.0081)**	**0.0965 (0.004)**	**0.6043 (0.0023)**	**3.158 (0.0248)**	NA

Since this data has a view that takes negative values, we do not compare against NMF-mvc

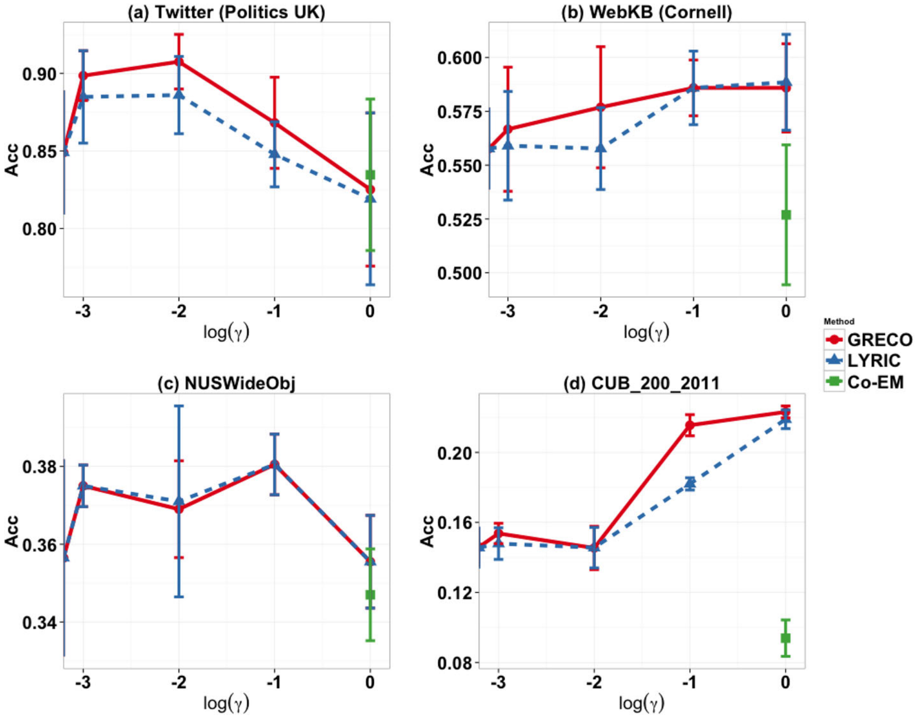

Fig. 3 Clustering accuracy of GRECO and LYRIC w.r.t. $\log \gamma$ on **a** Twitter data, **b** WebKB data, **c** NUSWideObj data and **d** CUB_200_2011 data

For Twitter data, the γ parameter of 0.01 resulted in the best clustering accuracy as measured on hold-out set (see Table 1). This provides further insight that the views do have some bias in the latent clustering distribution. In the absence of such a bias, the best clustering parameter should have corresponded to $\gamma \to 0$. Thus the value of the divergence parameter γ provides an intuitive understanding of inherent incoherence in clustering beliefs in the data. It is notable that characterizing this bias has resulted in almost an order of magnitude increase in clustering accuracy compared to baselines like multiview NMF and spectral clustering methods. To the best of our knowledge, there is little work in terms of designing robust learning models when underlying model assumptions may be violated. The results on Twitter data strongly highlight the significance of such an approach.

Similar observations on the WebKB data suggests a high degree of incoherence across views on the clustering distributions, suggested by the fact that linear aggregation ($\gamma \to 1$) provides the best results on the hold-out dataset. Note that in such a scenario, i.e. when views completely disagree (in terms of the MAP estimate of the clustering) across views, learning each view independently is equally useful, as demonstrated by competitive performance of Ensemble methods relative to GRECO and LYRIC. Again, this further reinforces the advantage of our model in terms of robustness to violations of model assumptions. Figure 3 also suggests that as the underlying bias is assumed increase, the model performance in both LYRIC and GRECO consistently improves. In addition, the improvement over Co-EM at $\gamma \to 1$ suggests that the method proposed to estimate a hold-out clustering assignment using (10) is better or comparable to that of Co-EM. Note that although GRECO and LYRIC do not perform the best on training data in terms of NMI and Entropy, the results on hold-out set are competitive—suggesting that the models do not overfit the training data.

From the results of NUS-Wide Object dataset, where six views are modeled jointly, the improvement in performance is significant when an appropriate divergence parameter γ is used, as compared to Co-EM, which enforces linear aggregation and the joint model that estimates a single clustering posterior across all views. This further suggests advantages of GRECO and LYRIC when the number of views available is large. The best performing divergence parameter is relatively high ($\gamma = 0.1$). This also suggests that as the number of views being modeled increases, the views are likely to be more incoherent and an assumption of a high bias (higher γ) is a better modeling assumption. This is also apparent from the deteriorated performance of the joint model. Both GRECO and LYRIC perform the best at the limiting case $w_g = 1$ as expected in a slightly high bias case, when additional regularization of GRECO is not necessarily advantageous. Figure 3 also suggests that at lower values of γ both LYRIC and GRECO may be getting stuck in local minima (suggested by the high observed variance at $\gamma = 0.01$) potentially reflecting sensitivity to choice of γ for this data.

For a large dataset like CUB-200-2011 with 200 clusters and ~11,000 samples and high dimensionality (~8000), the improvement in unsupervised learning performance of GRECO and LYRIC is more pronounced compared to Co-EM even though the best performance is obtained at $\gamma \rightarrow 1$. This suggests that our inference on hold-out set works better than Co-EM (see Sect. 4.6 for details). Further, the best performance divergence parameter $\gamma \rightarrow 1$ suggests the attribute view and the Fisher vector views, used from the CUB_200_2011 data, are potentially incoherent in terms of the latent clustering distribution. Comparison to other probabilistic methods, i.e. Joint model and Ensemble model, suggest restrictive model assumptions may fail and general methods like GRECO and LYRIC may be more reliable in large scale settings. Ensemble model also relies on Hungarian matching to solve the correspondence problem between cluster indices (200 clusters) across views. Improved performance in GRECO and LYRIC is obtained at a significant computational cost compared to CCA-mvc which provides comparable performance very fast. This corroborates the model assumptions made by CCA-mvc, namely that views of a sample are uncorrelated conditioned on cluster identity of sample (weaker assumptions than those made by the Joint model) can provide improvement in unsupervised learning performance. Faster inference for GRECO/LYRIC in such settings can be obtained by parallelization and/or any improvements to the variational inference procedure used to impose co-regularization.

Overall, the best Rényi divergence suitable for a particular dataset differs, indicating that GRECO and LYRIC capture potential differences in coherence between views with respect to cluster memberships significantly better than comparable methods. The biases between views demonstrably affect clustering performance. This also suggests that the multiview assumption of a single underlying cluster membership distribution is not always satisfied in real data. Thus flexible models such as GRECO and LYRIC are preferable. All results further show that the choice of the class of Rényi divergences is beneficial for improving multiview clustering performance and both methods generalize better to unseen data compared to baselines.

A comparison of training time suggests that the increased accuracy of GRECO and LYRIC is obtained at the cost of increased training time. However, as suggested in Sect. 4, the variational update required for co-regularization is the major contributing factor to training time. Since these updates can be trivially executed in a distributed setting across samples as well as for estimating unnormalized cluster membership distributions, the training time can be easily improved. Further, any alternative inference procedure to solve the co-regularization constraint will directly improve training times for the proposed method. Also note that training times are comparable to Co-EM and other baselines for special cases (see Tables 2, 4).

Additional advantages of GRECO and LYRIC compared to other methods are noteworthy. Both Twitter and WebKB datasets consist of at least one view with relational data. The twitter data is sparse (as is the case with social network data), i.e., a lot of the entries are 0. In these cases probabilistic methods outperform other methods suggesting the importance of probabilistic models in general. The NUS-Wide Object dataset and CUB datasets have mixed views, i.e. *bag-of-words* as well as numeric features (e.g. Fisher vector representations). Empirical evaluation also demonstrates that our methods handle mixed data well.

Some limitations of the proposed methods arise in selecting an appropriate choice of weights and the best suited Rényi divergence parameter for a given dataset. Storkey et al. (2014) have proposed a method for automatic selection of weights which can be easily incorporated in GRECO or LYRIC via minor changes to the variational procedures described in Appendix 1. However, we chose to use manual selection of weights inorder to highlight significance of the choice of Rényi divergences as opposed to a finer choice of weights, especially to highlight the generalization over Co-EM. Note that automatic selection or learning the best divergence parameter in an unsupervised setting suitable for a given data is a challenging and novel problem that we expose. Particularly, conventional model selection methods that trade-off model complexity and likelihood are not applicable in this scenario as model complexity does not change w.r.t. different γ. Automatic selection of such a model parameter is deferred to future work. However, we point out that both GRECO and LYRIC provide better performance compared to all existing baselines for all choices of γ that we tested. A more appropriate choice of γ further boosts performance. In case computational constraints exist, we suggest using either of the closed form methods suggested in Sect. 4.

6 Discussion and conclusion

This work proposed a co-regularization approach to multiview clustering that builds on a novel idea of directly minimizing a weighted sum of divergences between view-specific posteriors that indicate probabilities of cluster memberships. This approach encourages coherence between the posterior memberships by bringing them ‘closer’ in distribution. The resulting co-regularization techniques, GRECO and LYRIC significantly improve performance over existing multiview clustering methods. By maintaining per-view posteriors and using a flexible choice of Rényi divergences for imposing coherence, these models are robust to incoherence among views. In addition, Co-EM is recovered as a special case of LYRIC. Co-EM proposes linear aggregation of posteriors, which is best suited when aggregating among incoherent posterior memberships. We show empirically that better performance can be achieved by accounting for incoherence via a flexible family of divergences. We also achieve closed form updates to impose co-regularization for two special cases, when the divergence parameter $\gamma \to 0$ and $\gamma \to 1$.

For future work, a more general framework for multiview parameter estimation that accounts for divergence aggregation can be explored. Additional performance and computational gains may be obtained by learning the regularization weights and the divergence parameter γ. Theoretical analyses of special cases and studying the effects of other class of divergence can provide insights in further developing such flexible models. Such a framework could also offer advantages when views may be arbitrarily missing or in distributed settings when minimal interaction between views is expected due to communication constraints.

Acknowledgments This work is supported in part by NSF SCH-1417697 and the United States Office of Naval Research, Grant No. N00014-14-1-0039. MDR is supported by an ARC Discovery Early Career

Research Award (DE130101605). Authors would also like to acknowledge Suriya Gunasekar for reviewing the manuscript for clarity.

Appendix 1: Derivation of variational inference for weighted sum of divergence minimization

Minimize the weighted sum of Rényi divergences between M distributions $p^i(\mathbf{z}),\ i \in [M]$, given the divergence parameter γ. Let $q^*(\mathbf{z})$ be the corresponding minimizing distribution and $w \in \Delta^M$ be the known weight vector determining how important a given distribution is. The specific cost function is given by (13). Consider the case when each of the distributions are categorical distributions over clusters $[K]$.

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \sum_{i \in [M]} w_i \mathcal{D}_\gamma(p^i(\mathbf{z}) \| q(\mathbf{z})) \tag{13}$$

Let $\kappa^i(\mathbf{z})$ be a variational distribution corresponding to $p^i(\mathbf{z})$. Using the log-sum inequality, we have a lower bound on (13) given by (14).

$$\begin{aligned}&\sum_{i \in [M]} \frac{w_i}{\gamma} \mathcal{D}_\gamma(p^i(\mathbf{z}) || q(\mathbf{z})) \geq \\ &\sum_{i \in [M]} \frac{w_i}{\gamma(\gamma-1)} \sum_{k \in K} \kappa^i(z_k = 1)(\log[p^i(\mathbf{z})^\gamma q(\mathbf{z})^{(1-\gamma)}] - \log \kappa^i(z_k = 1))\end{aligned} \tag{14}$$

Algorithm 3 Variational Update to solve (14)

```
Given $\mathbf{w}$, $\gamma$ and initial estimates of $q(\mathbf{z})$,
repeat
  $\kappa^i(\mathbf{z}) \propto p^i(\mathbf{z})^\gamma q(\mathbf{z})^{(1-\gamma)}\ \forall i \in [M]$
  $q(\mathbf{z}) \propto \sum_{i \in [M]} w_i \kappa^i(\mathbf{z})$
until converged
```

The lower bound is optimized by iteratively estimating $\kappa^i(\mathbf{z})$'s and $q(\mathbf{z})$. To update $\kappa^i(\mathbf{z})$, $\kappa^j(\mathbf{z})\ \forall j \in [M],\ j \neq i$ and $q(\mathbf{z})$ are held fixed. Setting the gradient w.r.t. $\kappa^i(\mathbf{z})$, the iterative update is given by $\kappa^i(\mathbf{z}) \propto p^i(\mathbf{z})^\gamma q(\mathbf{z})^{(1-\gamma)}$. When all $\kappa^i(\mathbf{z})$ are held fixed, $q(\mathbf{z})$ is again obtained by setting the gradient of the bound w.r.t. $q(\mathbf{z})$ to 0, and given by (15). The complete variational update is described by Algorithm 3. Note that all distributions should be appropriately renormalized.

$$q(\mathbf{z}) \propto \sum_{i \in [M]} w_i \kappa^i(\mathbf{z}_n) \tag{15}$$

Algorithm 4 Variational Update to solve (3)

```
Given $\mathbf{w} \in \Delta^V$, $\gamma$ and $\boldsymbol{\Psi}_t^i\ \forall i \in [V]$
repeat
  $\kappa^i(\mathbf{z}_n) \propto p(\mathbf{z}_n^i | \mathbf{x}_n^i, \boldsymbol{\Psi}_t^i)^\gamma g(\mathbf{z}_n)^{(1-\gamma)}\ \forall i \in [V]$
  $g(\mathbf{z}_n) \propto \sum_{i \in [V]} w_i \kappa^i(\mathbf{z}_n)$
until converged
```

Algorithm 5 Variational Update to solve (4)

```
Given $w_g,\ w_v, \gamma, g^*(\mathbf{z}_n)$ and current parameter estimates, $\boldsymbol{\Psi}_t^v$
repeat
  $\kappa^*(\mathbf{z}_n) \propto g^*(\mathbf{z}_n)^\gamma q(\mathbf{z}_n^v)^{(1-\gamma)}$
  $\kappa^v(\mathbf{z}_n^v) \propto p(\mathbf{z}_n^v|\mathbf{x}_n^v, \boldsymbol{\Psi}_t^v)^\gamma q(\mathbf{z}_n^v)^{(1-\gamma)}$
  $q(\mathbf{z}_n^v) \propto w_g\kappa^*(\mathbf{z}_n) + w_v\kappa^v(\mathbf{z}_n^v)$
until converged
```

Appropriate variants of Algorithm 3 are used by GRECO and LYRIC. To estimate the centroid distribution of GRECO, Algorithm 4 is used. To estimate view-specific distributions $q_t(\mathbf{z}^v)\forall v \in [V]$, i.e. (4), Algorithm 5 is used. In the case of LYRIC, Algorithm 4 is used except in that the target distribution is $q_t(\mathbf{z}^v)$.

All the proposed variational updates (3) can be run in parallel for each sample $n \in [N]$. Further, for each sample, calculation of $\kappa^i(\mathbf{z})$ for each $i \in [M]$ and each $k \in [K]$ can be estimated in parallel up to proportionality. Similarly for the target variable $q(\mathbf{z})$, the estimates are trivially parallelizable for each $k \in [K]$ up to proportionality.

Appendix 2: Detailed derivation of parallel (co-regularization) E-step over samples, for GRECO and LYRIC

Let $\mathbf{z}^i = \{\mathbf{z}_n^i : n \in [N]\}$ and $\mathbf{x}^i = \{\mathbf{x}_n^i : n \in [N]\}$. Let $\mathbf{z} = \{\mathbf{z}^i : i \in [V]\}$, $\mathbf{x} = \{\mathbf{x}^i : i \in [V]\}$ and $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}^i : i \in [V]\}$. Let $g(\mathbf{z})$ be the target posterior for GRECO that is obtained by solving (16).

$$\begin{aligned} g(\mathbf{z}) &= \arg\min_{q(\mathbf{z})} \sum_{i\in[V]} w_i \mathcal{D}_\gamma(p(\mathbf{z}^i|\mathbf{x}^i, \boldsymbol{\Psi}^i)\|g(\mathbf{z})) \\ &= \arg\min_{q(\mathbf{z})} \sum_{i\in[V]} \frac{w_i}{\gamma(\gamma-1)} \log E_{p(\mathbf{z}|\mathbf{x},\boldsymbol{\Psi})}\left[\left(\frac{g(\mathbf{z})}{p(\mathbf{z}^i|\mathbf{x}^i, \boldsymbol{\Psi})}\right)^{(1-\gamma)}\right] \end{aligned} \tag{16}$$

We estimate $g(\mathbf{z})$ such that it is independent across all samples, i.e. $g(\mathbf{z}) = \prod_{n\in[N]} g(\mathbf{z}_n)$. By the IID assumption on the log-likelihood, the posterior $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\Psi})$ can be factored into per-view per-sample posteriors as in (17).

$$p(\mathbf{z}|\mathbf{x}, \boldsymbol{\Psi}) = \prod_{n\in[N]} \prod_{i\in[V]} p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i) \tag{17}$$

Therefore,

$$\begin{aligned} g(\mathbf{z}) &= \arg\min_{g(\mathbf{z})} \sum_{i\in[V]} w_i \log E_{\prod_{n\in[N]}\prod_{i\in[V]} p(\mathbf{z}_n^i|\mathbf{x}_n^i,\boldsymbol{\Psi}^i)}\left[\prod_{n\in[N]}\left(\frac{g(\mathbf{z}_n)}{p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i)}\right)^{(1-\gamma)}\right] \\ &= \arg\min_{g(\mathbf{z})} \sum_{i\in[V]} w_i \log \prod_{n\in[N]} E_{p(\mathbf{z}_n^i|\mathbf{x}_n^i,\boldsymbol{\Psi}^i)}\left[\left(\frac{g(\mathbf{z}_n)}{p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i)}\right)^{1-\gamma}\right] \\ &= \arg\min_{g(\mathbf{z})} \sum_{i\in[V]} w_i \sum_{n\in[N]} \log E_{p(\mathbf{z}_n^i|\mathbf{x}_n^i,\boldsymbol{\Psi}^i)}\left[\left(\frac{g(\mathbf{z}_n)}{p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i)}\right)^{1-\gamma}\right] \end{aligned}$$

$$= \arg\min_{\prod_{n\in[N]} g(\mathbf{z}_n)} \sum_{n\in[N]} \sum_{i\in[V]} w_i D_\gamma(p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i)\|g(\mathbf{z}_n))$$

$$\therefore g(\mathbf{z}_n) = \arg\min_{g(\mathbf{z}_n)} \sum_{i\in[V]} w_i D_\gamma(p(\mathbf{z}_n^i|\mathbf{x}_n^i, \boldsymbol{\Psi}^i)\|g(\mathbf{z}_n)) \tag{18}$$

Equation 18 can now be solved in parallel for each sample n to obtain $g(\mathbf{z}) = \prod_{n\in[N]} g(\mathbf{z}_n)$. This completes the proof and can be analogously proved for LYRIC and view-specific updates.

Appendix 3: Special cases of Rényi divergence aggregation

Case I : $\gamma \to 1$:

Storkey et al. (2014) have shown that weighted Rényi divergence aggregation when $\gamma \to 1$ is equivalent to (19)

$$q(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \sum_{i\in[V]} w_i \mathrm{KL}\Big(p(\mathbf{z}^i)\|q(\mathbf{z})\Big) \tag{19}$$

For multiview clustering, we aggregate between categorical distributions. Let $p(\mathbf{z}^i)$ be a categorical distribution parametrized by $\boldsymbol{\theta}^i$ so that $Pr(z_k^i = 1) = \theta_k^i$, $\boldsymbol{\theta}^i \in \Delta^K$. The target distribution $q(\mathbf{z})$, also categorical is parametrized by $\boldsymbol{\phi}$. Then the KL-divergence aggregation of (19) is given by (20)

$$\arg\min_{\boldsymbol{\phi}} \sum_{i\in[V]} w_i \Big\langle \boldsymbol{\theta}^i, \log\boldsymbol{\theta}^i - \log\boldsymbol{\phi}\Big\rangle \quad s.t. \boldsymbol{\phi} \in \Delta^K \tag{20}$$

(20) is convex in $\boldsymbol{\phi}$. The corresponding Lagrangian function is given by (21)

$$\arg\min_{\boldsymbol{\phi}} \sum_{i\in[V]} w_i \Big\langle \boldsymbol{\theta}^i, \log\boldsymbol{\phi} - \log\boldsymbol{\theta}^i\Big\rangle + \lambda\left(\sum_{k\in[K]} \phi_k - 1\right) + \beta^T\boldsymbol{\phi}, \ \beta \geq 0. \tag{21}$$

Setting the gradient of (21) to 0,

$$-\sum_{i\in[V]} w_i \frac{\boldsymbol{\theta}^i}{\boldsymbol{\phi}} + \lambda\mathbf{1} + \beta = 0 \tag{22}$$

If $\lambda\mathbf{1} + \beta = 1$, $\boldsymbol{\phi} = \sum_{i\in[V]} w_i\boldsymbol{\theta}^i$and $\boldsymbol{\phi} \in \Delta^K$, a feasible solution.

Case II : $\gamma \to 0$:

Storkey et al. (2014) have shown that weighted Rényi divergence aggregation when $\gamma \to 1$ is equivalent to (19)

$$q(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \sum_{i\in[V]} w_i \mathrm{KL}\Big(q(\mathbf{z})\|p(\mathbf{z}^i)\Big) \tag{23}$$

For multiview clustering, we aggregate between categorical distributions. Let $p(\mathbf{z}^i)$ be a categorical distribution parametrized by $\boldsymbol{\theta}^i$ so that $Pr(z_k^i = 1) = \theta_k^i$, $\boldsymbol{\theta}^i \in \Delta^K$. The target distribution $q(\mathbf{z})$, also categorical is parametrized by $\boldsymbol{\phi}$. Then the KL-divergence aggregation of (23) is given by (24)

$$\arg\min_{\boldsymbol{\phi}} \sum_{i\in[V]} w_i \Big\langle \boldsymbol{\phi}, \log\boldsymbol{\phi} - \log\boldsymbol{\theta}^i\Big\rangle \quad s.t. \boldsymbol{\phi} \in \Delta^K \tag{24}$$

(24) is convex in $\boldsymbol{\phi}$. The corresponding Lagrangian function is given by (25)

$$\arg\min_{\boldsymbol{\phi}} \sum_{i\in[V]} w_i \left\langle \boldsymbol{\phi}, \log\boldsymbol{\phi} - \log\boldsymbol{\theta}^i \right\rangle + \lambda \left(\sum_{k\in[K]} \phi_k - 1 \right) + \beta^T \boldsymbol{\phi},\ \beta \geq 0. \tag{25}$$

Setting the gradient of (25) to 0 as before, we have,

$$(1 + \log\boldsymbol{\phi}) - \sum_{i\in[V]} w_i \log\boldsymbol{\theta}^i + \lambda\mathbf{1} + \beta = 0 \tag{26}$$

If $\lambda\mathbf{1} + \beta + 1 = 0$, $\log\boldsymbol{\phi} = \sum_{i\in[V]} w_i \log\boldsymbol{\theta}^i$and $\boldsymbol{\phi} \in \Delta^K$.

Appendix 4: M-step for standard mixture models

Let N be the total number of samples in a mixture model with K classes. Let at any iteration $t, q(\mathbf{z}_n)$ be the posterior responsibilities calculated using current model parameters of the mixture model. Let $\mathbf{x}_n \in \mathbb{R}^D$ represent the observed features e.g. numeric data modeled as a Gaussian mixture or *bag-of-words* data that can be modeled as a mixture of multinomials.

- **Gaussian mixture models:** If the mixture model is a Gaussian mixture with parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k \forall k \in [K]$, the mean $\boldsymbol{\mu}_k$ and Covariance $\boldsymbol{\Sigma}_k$ are updated using (27) and (28) respectively.

$$\boldsymbol{\mu}_{t+1,k} = \frac{\sum_{n\in[N]} q(\mathbf{z}_{n,k})\mathbf{x}_n}{\sum_{n\in[N]} q(\mathbf{z}_{n,k})} \tag{27}$$

$$\boldsymbol{\Sigma}_{t+1,k} = \frac{\sum_{n\in[N]} q(\mathbf{z}_{n,k})(\mathbf{x}_n - \boldsymbol{\mu}_{t+1,k})(\mathbf{x}_n - \boldsymbol{\mu}_{t+1,k})^T}{\sum_{n\in[N]} q(\mathbf{z}_{n,k})} \tag{28}$$

- **Multinomial mixture models:** The Multinomial distribution parameters for each cluster $\theta_k \ \forall k \in [K]$ can be updated using (29)

$$\boldsymbol{\theta}_{t+1,k} = \frac{\sum_{n\in[N]} q(\mathbf{z}_{n,k})\mathbf{x}_n}{\sum_{n\in[N]} q(\mathbf{z}_{n,k}) \sum_{d\in[D]} \mathbf{x}_{n,d}} \tag{29}$$

Appendix 5: Formulae of evaluation metrics

All evaluation metrics assume that ground-truth cluster memberships are known and that the correspondence between clustering labels and ground-truth labels is estimated. The number of learned clusters is the same as number of ground-truth clusters.

Definition 2 If C_n represents the cluster label determined by the learning algorithm and ω_n represents the ground-truth clustering, the **clustering accuracy** for a dataset with N samples and K clusters is given by,

$$\textbf{Accuracy} = \frac{\sum_{n\in[K]} \sum_{k\in[K]} \mathbf{1}(C_n == \omega_n)}{N}$$

where,

$$\mathbf{1}(C_n == \omega_n) = \begin{cases} 1, & \text{if } C_n = \omega_n, \\ 0, & \text{otherwise.} \end{cases}$$

Following terms are defined per cluster $k \in [K]$

- *True Positives* (TP_k): It is the number of samples that were clustered correctly by the learning model.
- *False Positives* (FP_k): It is the number of samples assigned to a cluster they do not belong to.
- *True Negatives* (TN_k): This is the total number of samples not belonging to a given cluster and is clustered correctly i.e. clustered into a different cluster than for which true negatives are measured.
- *False Negatives* (FN_k): This is the total number of samples belonging to a given cluster that were not actually assigned to the cluster by the learning algorithm.

Definition 3

$$\textbf{Precision} = \frac{\sum_{k\in[K]} TP_k}{\sum_{k\in[K]} TP_k + FP_k}$$

Definition 4

$$\textbf{Recall} = \frac{\sum_{k\in[K]} TP_k}{\sum_{k\in[K]} TP_k + FN_k}$$

Definition 5

$$\textbf{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The following metrics do not assume a correspondence between ground-truth labels and learned cluster labels.

Definition 6 Let C be the categorical random variable over K clusters with a distribution obtained from clustering i.e. $Pr(C = k)$ is the fraction of samples clustered into k by the learning algorithm. Let ω represent the categorical variable with a distribution obtained from true clustering. The joint distribution $p(C, \omega)$ is the fraction of samples clustered as C and lie in ground-truth cluster ω. The mutual information $I(C, \omega)$ is given by,

$$I(C,\omega) = \sum_{k\in[K]} \sum_{j\in[K]} p(C = k, \omega = j) \log \frac{p(C = k, \omega = j)}{p(C = k)p(\omega = j)}$$

The Entropy of $H(C) = -\sum_{k\in[K]} p(C = k) \log p(C = k)$ and analogously for $H(\omega)$. **Normalized Mutual Information (NMI)** (Strehl and Ghosh 2003) is the symmetrized and normalized mutual information between C and ω.

$$NMI(C,\omega) = \frac{I(C,\omega)}{\frac{H(C)+H(\omega)}{2}}$$

Definition 7 **Average Entropy** (Bickel and Scheffer 2005)

$$= -\sum_{j\in[K]} p(C = j) \sum_{k\in[K]} p(C = j, \omega = k) \log p(C = j, \omega = k)$$

Appendix 6: Empirical convergence of log-likelihood

In order to validate empirical convergence in log-likelihood for GRECO and LYRIC for a fixed divergence parameter γ and the co-regularizing weights, we conduct the following

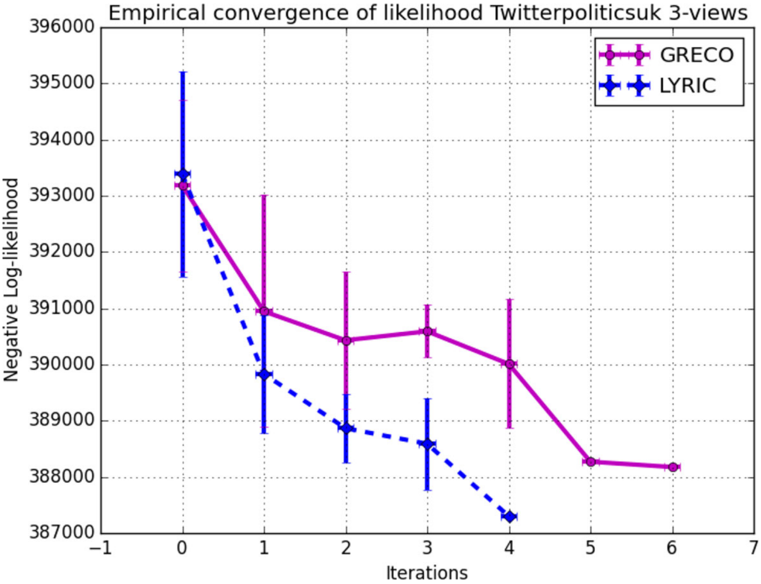

Fig. 4 Twitter-politicsuk dataset

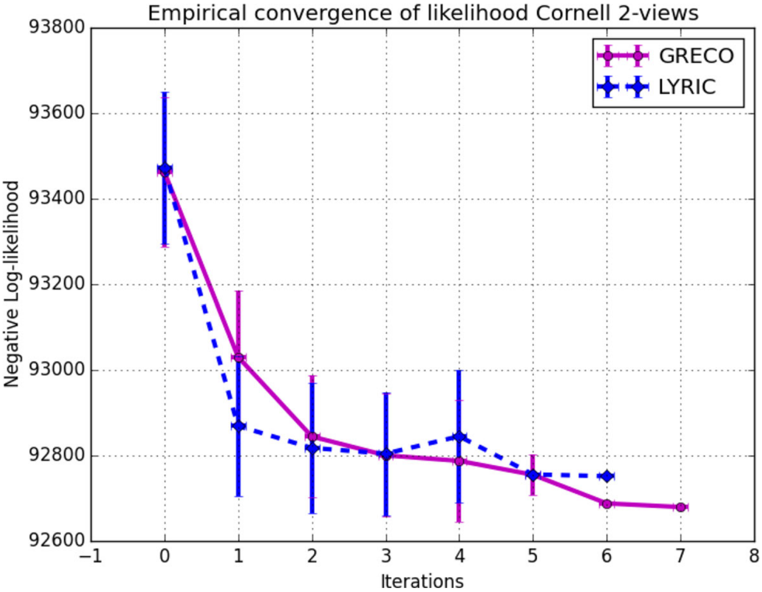

Fig. 5 Cornell (WebKB) dataset

experiment. For all datasets, a single train-test split is chosen and the parameters are initialized differently (using k-means clustering) for five trials and the negative log-likelihood (- $\sum_{v\in[V]}\sum_{n\in[N]}\log p(\mathbf{x}_n^v, \mathbf{z}_n^v; \boldsymbol{\Psi}_n^v)$) is observed over iterations until convergence. Figures 4, 5, 6 and 7 show the negative log-likelihood observed for each alternating EM iteration for

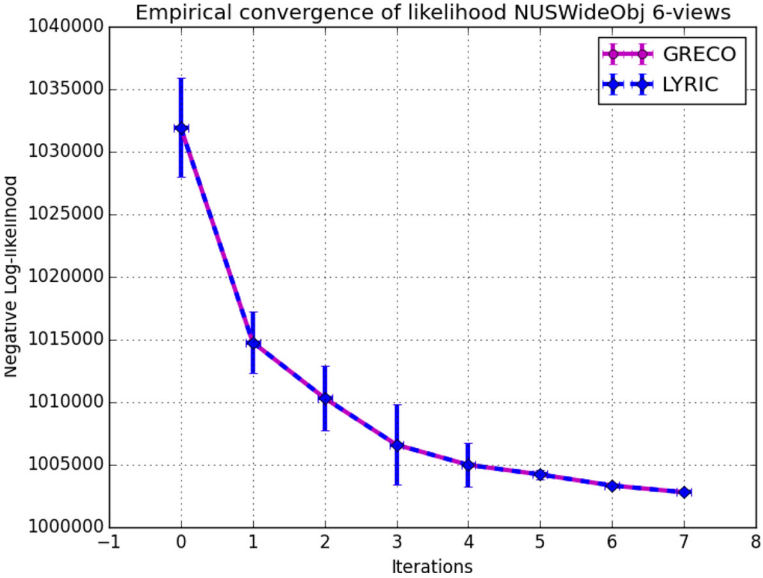

Fig. 6 NUSWideObj dataset

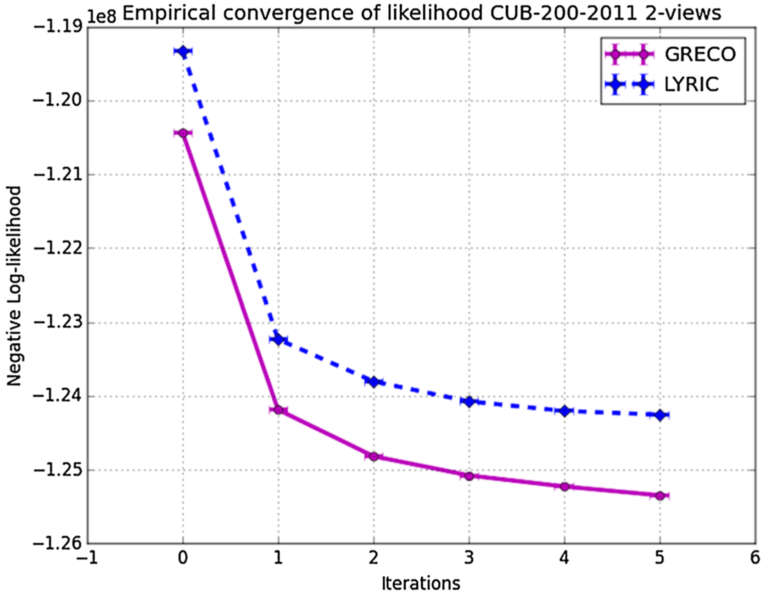

Fig. 7 CUB-200-2011 dataset

both GRECO and LYRIC for the Twitter-politicsuk, Cornell (WebKB), NUSWideObj and CUB-200-2011 datasets respectively.

References

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2013). Label-embedding for attribute-based classification. In *IEEE conference on computer vision and pattern recognition (CVPR), 2013*, pp. 819–826.

Akata, Z., Thurau, C., & Bauckhage, C. (2011). Non-negative matrix factorization in multimodality data for segmentation and label prediction. In: W. Andreas, S. Sabine & G. Martin (Eds.), *16th Computer vision winter workshop*.

Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161–163.

Bickel, S., & Scheffer, T. (2005). Estimation of mixture models using Co-EM. In *Machine learning: ECML 2005, 16th European conference on machine learning, Porto, Portugal*.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on computational learning theory*. ACM.

Chaudhuri, K., Kakade, S. M., Livescu, K., & Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y.-T. (2009). NUS-WIDE: A real-world web image database from National University of Singapore. In *Proceedings of ACM conference on image and video retrieval*.

Dasgupta, S., Littman, M. L., & McAllester, D. A. (2001). PAC generalization bounds for co-training. In *Advances in neural information processing systems NIPS*.

de Sa V. R., & Ballard, D. H. (1993). *Self-teaching through correlated input*. In Computation and neural systems. New York: Springer.

De Sa, V. R. (2005). Spectral clustering with two views. In *Proceedings of the workshop on learning with multiple views, international conference on machine learning*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, *39*, 1–38.

Eaton, E., des Jardins, M., & Jacob, S. (2014). Multi-view constrained clustering with an incomplete mapping between views. *Knowledge and Information Systems*, *38*, 231–257.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., & Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129.

Garg, A., Jayram, T. S., Vaithyanathan, S., & Zhu, H. (2004). Generalized opinion pooling. In *AMAI*.

Geoffrey, E. (2002). Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14* (8), 1771–1800.

Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 305–315.

Greene, D., & Cunningham, P. (2013). Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th annual ACM web science conference*.

Guo, Y. (2013). Convex subspace representation learning from multi-view data. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence, 2013*.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2* (1), 83–97.

Kumar, A., & Daume-III, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning*.

Kumar, A., Rai, P., & Daumé III, H. (2010). Co-regularized spectral clustering with multiple kernels. In *NIPS Workshop on new directions in multiple kernel learning*.

Kumar, A., Rai, P., & Daume, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems 24*. Curran Associates, Inc.,.

Li, S.-Y., Jiang, Y., & Zhou, Z.-H. (2014). Partial multi-view clustering. In *AAAI Conference on artificial intelligence*.

Lian, W., Rai, P., Salazar, E., & Carin, L. (2015). Integrating features and similarities: Flexible models for heterogeneous multiview data. In *Twenty-ninth AAAI conference on artificial intelligence*.

Liu, J., Wang, C., Gao, J., & Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of 2013 SIAM data mining conference*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on information and knowledge management*. ACM.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the 4th Berkeley symposium on mathematics, statistics and probability*.

Schmidhuber, J., & Prelinger, D. (1993). A novel unsupervised classification method. In *Third international conference on artificial neural networks, 1993*.

Sindhwani, V., & Rosenberg, D. S. (2008). An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on machine learning*. ACM.

Storkey, A., Zhu, Z., & Hu, J. (2014). A continuum from mixtures to products: Aggregation under bias. In *ICML workshop on divergence methods for probabilistic inference*.

Strehl, A., & Ghosh, J. (2003). Cluster ensembles–A knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, *3*, 583–617.

Tzortzis, G., & Likas, A. (2012). Kernel-based weighted multi-view clustering. *2013 IEEE 13th international conference on data mining*.

van Erven, T., & Harremoës, P. (2012). Rényi divergence and Kullback-Leibler divergence. *ArXiv e-prints*.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Weston, J., Bengio, S., & Usunier, N. (2011). Wsabie: Scaling up to large vocabulary image annotation. *In IJCAI*, (Vol. 11, pp. 2764–2770).

Zhou, D., & Burges, C. J. C. (2007). Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on machine learning*.