A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation

Arindam Banerjee Inderjit Dhillon
Joydeep Ghosh Srujana Merugu
University of Texas
Austin, TX, USA

Dharmendra Modha
IBM Almaden Research Center
San Jose, CA, USA

ABSTRACT

Co-clustering, or simultaneous clustering of the rows and columns of two-dimensional data matrices, is a powerful data mining technique with varied applications such as text clustering, microarray analysis and recommender systems. An information-theoretic approach that is applicable when the data matrix can be interpreted as a two-dimensional empirical joint probability distribution, was recently proposed. However, in many situations, co-clustering of more general matrices is desired. In this paper, we present a substantially generalized co-clustering framework wherein (i) loss functions corresponding to all Bregman divergences, which include squared Euclidean distance and KL-divergence as special cases, can be used, thereby making it applicable to a wide range of data matrices, (ii) various conditional expectation based constraints can be considered based on the statistics that need to be preserved, thereby giving rise to different parametric co-clustering models, and (iii) the maximum entropy principle is generalized to the minimum Bregman information principle to provide a natural model selection technique. The analysis yields an elegant meta algorithm that is guaranteed to achieve local optimality. Our methodology encompasses a vast majority of previously known clustering and co-clustering algorithms based on alternate minimization. We provide examples and empirical evidence to establish the generality and efficacy of the proposed co-clustering framework.

1. INTRODUCTION

Co-clustering, or bi-clustering [10, 5], is the problem of simultaneously clustering rows and columns of a data matrix. The problem of co-clustering arises in diverse data mining applications, such as simultaneous clustering of genes and experimental conditions in bioinformatics [5, 6], documents and words in text mining [9], users and movies in recommender systems, etc. Often, it forms a key intermediate step in the data mining process and is essential to overcome the noise and sparsity in the input data matrix. Further, co-clustering is capable of providing compressed representations that are highly interpretable while preserving most of the information contained in the original data, which makes it valuable to a large class of statistical data analysis applications.

In order to design a co-clustering framework, we need to first characterize the "goodness" of a co-clustering. Existing co-clustering techniques [6, 5, 9] achieve this by quantifying the "goodness" of a co-clustering in terms of the approximation error between the original data matrix and a matrix reconstructed by co-clustering based on the summary statistics. Currently, the most efficient and scalable ones are those based on alternate minimization schemes [6, 9, 6] that allow only two distortion measures namely, KL-divergence and the squared Euclidean distance. Further, they also allow only a few matrix reconstruction schemes that involve preserving particular summary statistics of the original matrix. These two limitations restrict the applicability of these techniques to a small range of data matrices.

In this paper, we address the following two questions: *(a) what class of distortion functions admit efficient co-clustering algorithms based on alternate minimization?*, and *(b) what are the different possible matrix reconstruction schemes for these co-clustering algorithms?*. We show that alternate minimization based co-clustering algorithms work for a large class of distortion measures called Bregman divergences [1], which include squared Euclidean distance, KL-divergence, Itakura-Saito distance, etc., as special cases. Further, we demonstrate that for a given co-clustering, a large variety of approximation models are possible based on the type of summary statistics that need to be preserved. To achieve these results, we propose and use a new *minimum Bregman information principle* that simultaneously generalizes the maximum entropy and the least squares principles. Based on the proposed principle, and other related results, we develop an elegant meta-algorithm for the Bregman co-clustering problem with a number of desirable properties such as scalability and applicability to a wide range of data matrices. Most previously known parametric clustering and co-clustering algorithms based on alternative minimization follow as special cases of our methodology.

2. MOTIVATION

We start by reviewing information-theoretic co-clustering [9] and concretely motivating the need for a more general co-clustering framework.

Let X and Y be discrete random variables that take values in the sets $\{x_u\}[u]_1^m$ where $[u]_1^m$ denotes an index u running over $\{1, \cdots, m\}$ and $\{y_v\}[v]_1^n$ respectively. Suppose we are in the idealized situation where the joint probability distribution $p(X, Y)$ is known. In practice, p may be estimated from a contingency table or co-occurrence matrix. Suppose we want to co-cluster, or, simultaneously cluster X into k disjoint (row) clusters $\{\hat{x}_g\}$ $[g]_1^k$ and Y into l disjoint (column) clusters, $\{\hat{y}_h\}$ $[h]_1^l$. Let $\hat{X}$ and $\hat{Y}$ denote the corresponding clustered random variables that range over these sets. An information theoretic formulation of finding the optimal co-clustering is to solve the problem

$$\min_{\hat{X}, \hat{Y}} \quad I(X;Y) - I(\hat{X};\hat{Y}) \ , \tag{2.1}$$

where $I(X;Y)$ is the mutual information between X and Y [7]. In [9], it was shown that

$$I(X;Y) - I(\hat{X},\hat{Y}) = D(p(X,Y)||q(X,Y)), \tag{2.2}$$

where $q(X,Y)$ is a distribution of the form

$$q(X,Y) = p(\hat{X},\hat{Y})p(X|\hat{X})p(Y|\hat{Y}), \tag{2.3}$$

and $D(\cdot||\cdot)$ denotes the Kullback-Leibler(KL) divergence, also known as relative entropy. Thus, the search for the optimal co-clustering may be conducted by searching for the nearest approximation $q(X,Y)$ that has the above form. On closer examination, we note that the distribution $q(X,Y)$ depends only on $(kl + m + n - 3)$ independent parameters, which is much smaller than the $(mn - 1)$ parameters that determine a general joint distribution p. Hence, we call $q(X,Y)$ a "low complexity" or low parameter matrix approximation.

The above is the viewpoint presented in [9]. We now present an alternate viewpoint that will enable us to generalize our approach to arbitrary data matrices and general distortion measures. The following lemma highlights the key maximum entropy property that makes $q(X,Y)$ a "low complexity" or low parameter approximation.

Lemma 1 *Given a fixed co-clustering $\hat{X}$, $\hat{Y}$, consider the set of joint distributions p' that preserve the following statistics of the input distribution p:*

$$\sum_{x\in\hat{x}}\sum_{y\in\hat{y}} p'(x,y) = p(\hat{x},\hat{y}) = \sum_{x\in\hat{x}}\sum_{y\in\hat{y}} p(x,y), \quad \forall \hat{x},\hat{y},$$
$$p'(x) = p(x), \quad p'(y) = p(y), \quad \forall x,y.$$

Among all such distributions p', the distribution q in (2.3) has the maximum entropy, i.e.,

$$H(q(X,Y)) \geq H(p'(X,Y)).$$

PROOF. It can be easily checked that q preserves the relevant statistics so that $p'(\hat{X},\hat{Y}) = p(\hat{X},\hat{Y}) = q(\hat{X},\hat{Y})$, $p'(X) = p(X) = q(X)$ and $p'(Y) = p(Y) = q(Y)$. Using this property of q, it is easy to show that $H(q) - H(p') = D(p'||q) \geq 0$. □

What is the significance of the above lemma? In the absence of any constraints, the uniform distribution, $p_0(X,Y) = \{\frac{1}{mn}\}$, has the maximum entropy. If only row and column marginals are to be preserved, then the product distribution $p(X)p(Y)$ has maximum entropy (see [7, Problem 5, Chap. 11]). The above lemma states that among all distributions that preserve marginals as well as co-cluster statistics, the maximum entropy distribution has the form in (2.3). It is important to note that this *maximum entropy characterization is equivalent to saying that q is a low-complexity matrix approximation.* Thus, by (2.2) and Lemma 1, *the co-clustering problem (2.1) is equivalent to the problem of finding the nearest (in KL-divergence) maximum entropy distribution that preserves the marginals, and the co-cluster statistics of the original data matrix.*

The above formulation is applicable when the data matrix corresponds to an empirical joint distribution. However, there are important situations when the data matrix is more general, for example, the matrix may contain negative entries and/or a distortion measure other than KL-divergence, such as the squared Euclidean distance, might be more appropriate.

This paper addresses the general situation by extending the information-theoretic co-clustering along three different directions. First, "nearness" can be now measured by any one of a large class of distortion measures called Bregman divergences. Second, we allow specification of a larger variety of constraints that preserve various statistics of the data. The different constraints allow a trade-off between complexity and fidelity of the resulting approximation. Lastly, to accomplish the above, we generalize the maximum entropy approach: we guide our co-clustering generalization by appealing to the *minimum Bregman information principle* that we shall introduce shortly. *The optimal co-clustering is guided by the search for the nearest (in Bregman divergence) matrix approximation that has minimum Bregman information while satisfying the constraints mentioned above.*

3. FORMULATION AND ANALYSIS

In this section, we formulate the Bregman co-clustering problem in terms of the Bregman divergence between a given matrix and an approximation based on the co-clustering. We show that a natural way of specifying the approximation matrix leads to a new minimum Bregman information principle, which we analyze in detail.

3.1 Preliminaries

We start by defining Bregman divergences [1, 3]. Let ϕ be a real-valued strictly convex function defined on the convex set $S = \text{dom}(\phi) \subseteq \mathbb{R}$, the domain of ϕ, such that ϕ is differentiable on $\text{int}(S)$, the interior of S. The **Bregman divergence** $d_\phi : S \times \text{int}(S) \mapsto [0, \infty)$ is defined as $d_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - \langle z_1 - z_2, \nabla\phi(z_2)\rangle$, where $\nabla\phi$ is the gradient of ϕ.

Example 1.A (I-Divergence) Given $z \in \mathbb{R}_+$, let $\phi(z) = z \log z$. For $z_1, z_2 \in \mathbb{R}_+$, $d_\phi(z_1, z_2) = z_1 \log(z_1/z_2) - (z_1 - z_2)$.

Example 2.A (Squared Euclidean Distance) Given $z \in \mathbb{R}$, let $\phi(z) = z^2$. For $z_1, z_2 \in \mathbb{R}$, $d_\phi(z_1, z_2) = (z_1 - z_2)^2$.

Data Matrix

We focus on the problem of approximating a given $m \times n$ data matrix Z under various constraints. Let each entry of Z take values in a convex set $S = \text{dom}(\phi)$. Hence, Z takes values in $S^{m\times n}$. Observe that we are now admitting a much larger classes of matrices than that in [9, 6].

We will think of Z as a random variable that is a known deterministic function of two underlying random variables

U and V, which we now introduce. Let U be a random variable taking values in $\{1, \cdots, m\}$, the set of row indices, and let V be a random variable taking values in $\{1, \cdots, n\}$, the set of column indices. Hence, the matrix $Z = [z_{uv}]$ is such that z_{uv} is some fixed deterministic function of u and v. Let $\nu = \{\nu_{uv} : [u]_1^m, [v]_1^n\}$ denote the joint probability measure of the pair (U, V), which is either pre-specified or set to be the uniform distribution. Throughout the paper, all expectations are with respect to ν.

Example 1.B (I-Divergence) Let $(X, Y) \sim p(X, Y)$ be jointly distributed random variables with X, Y taking values in $\{x_u\}, [u]_1^m$ and $\{y_v\}, [v]_1^n$ respectively. Then, $p(X, Y)$ can be written in the form of the matrix $Z = [z_{uv}]$, $[u]_1^m, [v]_1^n$, where $z_{uv} = p(x_u, y_v)$ is a deterministic function of u and v. This example with a uniform measure ν corresponds to the setting described in section 2 (originally in [9])[1].

Example 2.B (Squared Euclidean Distance) Let $Z \in \mathbb{R}^{m \times n}$ denote a data matrix whose elements may assume positive, negative, or zero values and let ν be a uniform measure. This example corresponds to the setting described in [6, 5].

Bregman Co-clustering

We define a $k \times l$ co-clustering as a pair of maps:

$$\rho : \{1, \cdots, m\} \mapsto \{1, \cdots, k\}$$
$$\gamma : \{1, \cdots, n\} \mapsto \{1, \cdots, l\}.$$

Let $\hat{U}$ and $\hat{V}$ be random variables taking values in $\{1, \cdots, k\}$ and $\{1, \cdots, l\}$ such that $\hat{U} = \rho(U)$ and $\hat{V} = \gamma(V)$. Let $\hat{Z} = [\hat{z}_{uv}] \in S^{m \times n}$ be an approximation for the data matrix Z such that it depends upon a given co-clustering (ρ, γ). We shall then measure the goodness of the underlying co-clustering as:

$$E[d_\phi(Z, \hat{Z})] = \sum_{u=1}^{m} \sum_{v=1}^{n} \nu_{uv} d_\phi(z_{uv}, \hat{z}_{uv}). \qquad (3.4)$$

To carry out this plan, we need to make precise the connection between (ρ, γ) and $\hat{Z}$.

Example 1.C (I-Divergence) The Bregman co-clustering objective function in this case is given by $E[d_\phi(Z, \hat{Z})] = E[Z \log(Z/\hat{Z}) - Z + \hat{Z}]$.

Example 2.C (Squared Euclidean Distance) The Bregman co-clustering objective function in this case is given by $E[d_\phi(Z, \hat{Z})] = E[(Z - \hat{Z})^2]$

3.2 Co-Clustering and Matrix Approximation

Every co-clustering can lead to numerous different matrix approximations. The crucial point is precisely what information from Z do we retain.

Let us fix a co-clustering (ρ, γ). Given the co-clustering, there are essentially five random variables of interest: Z, U, V, $\hat{U}$, and $\hat{V}$. Now, we can specify the statistics of Z that we want to preserve using non-trivial combinations from this set, given by

$$\Gamma = \{\{U, \hat{V}\}, \{\hat{U}, V\}, \{\hat{U}, \hat{V}\}, \{U\}, \{V\}, \{\hat{U}\}, \{\hat{V}\}\},$$

where $\{U, V\}$ is not included since $E[Z|U, V] = Z$. We will be interested in random variables that depend on sets of conditional expectations[2] of the form $\{E[Z|C],\ C \in \Gamma\}$. If $\pi(\Gamma)$ denotes the power set of Γ, then every element of $\pi(\Gamma)$ is a set of constraints, and leads to a (possibly) different matrix approximation. Intuitively, we think of $\pi(\Gamma)$ as the *class of matrix approximation schemes* related to a given co-clustering (ρ, γ).

We now display four concrete examples of interesting elements of $\pi(\Gamma)$ that we will use throughout this paper to illuminate discussions:

$$\mathcal{C}_1 = \{\{\hat{U}\}, \{\hat{V}\}\}, \qquad \mathcal{C}_2 = \{\{\hat{U}, \hat{V}\}\}$$
$$\mathcal{C}_3 = \{\{\hat{U}, \hat{V}\}, \{U\}, \{V\}\} \quad \mathcal{C}_4 = \{\{U, \hat{V}\}, \{\hat{U}, V\}\}$$

The diligent reader may verify that these are the only non-trivial symmetric constraint sets in $\pi(\Gamma)$. Also, observe that if we have access to $\{E[Z|C] : C \in \mathcal{C}_i\}$, for some $1 \leq i \leq 4$, then we can compute $\{E[Z|C] : C \in \mathcal{C}_j\}$ for all $1 \leq j \leq i$. In this sense, we say that the constraint set $\mathcal{C}_i$ is more complex than $\mathcal{C}_j$ for all $j \leq i$. From a practical perspective, a more complex set of constraints allows us to retain more information about Z.

Our abstraction allows us to handle the essence behind the above constraint sets and, in fact, all constraint sets in $\pi(\Gamma)$ at once. Now, consider an element $\mathcal{C}$ in the power set $\pi(\Gamma)$ as the pertinent constraint set. Given this choice, we seek to find the "best" approximation matrix. Let $\Xi_A(\rho, \gamma, \mathcal{C})$ denote the class of random variables $Z' \in S^{m \times n}$ that satisfy the following *conditional independence* condition:

Condition A. The Markov condition

$$Z \rightarrow \{E[Z|C] : C \in \mathcal{C}\} \rightarrow Z'$$

holds. In other words, Z' depends upon Z only through the set of random variables $\{E[Z|C] : C \in \mathcal{C}\}$.

We define the "best" matrix approximation $\hat{Z}_A$ corresponding to the co-clustering (ρ, γ) and the constraint set $\mathcal{C}$ as the one in the class $\Xi_A(\rho, \gamma, \mathcal{C})$ that minimizes the approximation error, i.e.,

$$\hat{Z}_A \equiv \hat{Z}_A(\rho, \gamma, \mathcal{C}) = \underset{Z' \in \Xi_A(\rho, \gamma, \mathcal{C})}{\operatorname{argmin}} E[d_\phi(Z, Z')]. \qquad (3.5)$$

Before we proceed further, we will furnish an alternative characterization of (3.5) in terms of an extremely useful concept called Bregman information. This alternative characterization will be an important step in our hunt for a generalized algorithm.

3.3 Minimum Bregman Information

For any random variable Z', its **Bregman information** is defined as the expected Bregman divergence to the expectation, i.e.,

$$I_\phi(Z') = E[d_\phi(Z', E[Z'])]. \qquad (3.6)$$

Intuitively, this quantity is a measure of the "spread" or the "information" in the random variable.

Example 1.D (I-Divergence) Given $Z' \in \mathbb{R}_+^{m \times n}$, the Bregman information corresponding to I-divergence is given by

$$I_\phi(Z') = E[Z' \log(Z'/E[Z'])].$$

[1]Note that in [9] KL-divergence was used, which is a special case of I-divergence applicable to probability distributions.

[2]Given a sub-σ-algebra $\mathcal{G}$ for Z, the conditional expectation $E[Z|\mathcal{G}]$ is the optimal Bregman predictor among all $\mathcal{G}$-measurable random variables [2].

When Z' corresponds to a probability distribution, i.e., $z'_{uv} = p(x_u, y_u)$ and ν is a uniform measure, then $E[Z']$ is the uniform distribution p_0 and the Bregman information is given by $D(p||p_0) = -H(p)+$ constant, where $D(\cdot||\cdot)$ is KL-divergence and $H(.)$ is the Shannon entropy.

Example 2.D (Squared Euclidean Distance) Given $Z' \in \mathbb{R}^{m\times n}$, the Bregman information corresponding to squared Euclidean distance is given by $I_\phi(Z') = E[(Z' - E[Z'])^2]$, which is just the squared Frobenius norm of the matrix for a uniform measure ν.

We now consider a different class of approximating random variables based on a specified constraint set $\mathcal{C}$ and a specified co-clustering (ρ, γ). Let $\Xi_B(\rho, \gamma, \mathcal{C})$ denote a class of random variables such that every Z' satisfies the following *linear constraints*:

Condition B. For every $C \in \mathcal{C}$, $E[Z|C] = E[Z'|C]$.

With respect to the set $\Xi_B(\rho, \gamma, \mathcal{C})$, we ask: What is the "best" random variable to select from this set? We now propose a new **minimum Bregman information principle** that recommends selecting a random variable that has the minimum Bregman information subject to the linear constraints:

$$\hat{Z}_B \equiv \hat{Z}_B(\rho, \gamma, \mathcal{C}) = \underset{Z' \in \Xi_B(\rho,\gamma,\mathcal{C})}{\operatorname{argmin}} I_\phi(Z'). \tag{3.7}$$

It is easy to see that the widely used *maximum entropy principle* [12, 7] is a special case of the proposed principle since the entropy of a joint distribution is negatively related to the Bregman information (Example 1.D). In fact, the minimum Bregman information principle neatly unifies both the maximum entropy, and the least squares principle [8].

The following theorem characterizes the solution to the minimum Bregman information problem (3.7).

Theorem 1 *For a Bregman divergence d_ϕ, any random variable $Z \in S^{m\times n}$, a specified co-clustering (ρ, γ) and a specified constraint set $\mathcal{C}$, the solution $\hat{Z}_B$ to (3.7) is given by*[3]

$$\nabla\phi(\hat{Z}_B) = -\sum_{r=1}^{s} \Lambda_r^\star,$$

where $\Lambda^\star \equiv \{\Lambda_r^\star\}$ are the optimal Lagrange multipliers corresponding to set of the linear constraints:

$$E[Z'|C_r] = E[Z|C_r], \quad [r]_1^s.$$

Furthermore, $\hat{Z}_B$ always exists, is unique, and satisfies Condition A.

Proof. Consider the Lagrangian $J(Z', \Lambda)$ of the minimum Bregman information problem. After some algebraic manipulation, it can be shown that

$$\begin{aligned} J(Z',\Lambda) &= I_\phi(Z') + \sum_{r=1}^{s} \Lambda_r(E[Z'|C_r] - E[Z|C_r]) \\ &= E[\phi(Z')] - \phi(E[Z']) + \sum_{r=1}^{s} \Lambda_r(E[Z'|C_r] - E[Z|C_r]) \ . \end{aligned}$$

[3] In general, we use $f(Z)$ to denote $f(\cdot)$ applied to Z elementwise for any function f.

Table 1: Minimum Bregman information solution for I-Divergence.

Constraints $\mathcal{C}$	Approximation $\hat{Z}_B$
$\mathcal{C}_1$	$\frac{E[Z\|\hat{U}]\times E[Z\|\hat{V}]}{E[Z]}$
$\mathcal{C}_2$	$E[Z\|\hat{U},\hat{V}]$
$\mathcal{C}_3$	$\frac{E[Z\|U]\times E[Z\|V]\times E[Z\|\hat{U},\hat{V}]}{E[Z\|\hat{U}]\times E[Z\|\hat{V}]}$
$\mathcal{C}_4$	$\frac{E[Z\|U,\hat{V}]\times E[Z\|\hat{U},V]}{E[Z\|\hat{U},\hat{V}]}$

Table 2: Minimum Bregman information solution for squared Euclidean distance.

Constraints $\mathcal{C}$	Approximation $\hat{Z}_B$
$\mathcal{C}_1$	$E[Z\|\hat{U}] + E[Z\|\hat{V}] - E[Z]$
$\mathcal{C}_2$	$E[Z\|\hat{U},\hat{V}]$
$\mathcal{C}_3$	$E[Z\|U] + E[Z\|V] + E[Z\|\hat{U},\hat{V}]$ $-E[Z\|\hat{U}] - E[Z\|\hat{V}]$
$\mathcal{C}_4$	$E[Z\|U,\hat{V}] + E[Z\|\hat{U},V] - E[Z\|\hat{U},\hat{V}]$

Now, the Lagrange dual, $L(\Lambda) = \inf_{Z'} J(Z', \Lambda)$, is strictly concave in Λ. By maximizing the Lagrange dual we get the optimal Lagrange multipliers, i.e., $\Lambda^* = \{\Lambda_r^*\} = \operatorname{argmax}_\Lambda L(\Lambda)$. Replacing Λ^* into the first order necessary conditions corresponding to the minimizer $\hat{Z}_B$, we get

$$\nabla J(\hat{Z}_B, \Lambda^*) = 0 \quad \Leftrightarrow \quad \nabla\phi(\hat{Z}_B) + \sum_{r=1}^{s} \Lambda_r^* = 0 \ .$$

Rearranging terms proves the first part of the theorem.

The existence and the uniqueness of $\hat{Z}_B$ follow from the strict convexity of ϕ. Also, observe that due to the fact that the minimization problem takes as input only the set $\{E[Z|C] : C \in \mathcal{C}\}$, and, hence, has no other information about Z, it follows that $\hat{Z}_B$ satisfies Condition A. □

Example 1.E (I-Divergence) When $\phi(z) = z\log z, z \in \mathbb{R}_+$, $\nabla\phi(z) = \log z$ and the minimum Bregman information solution is given by $\log \hat{Z}_B = -\sum_{r=1}^{s} \Lambda_r^*$ where $\Lambda^\star = \{\Lambda_r^\star\}$ are the optimal Lagrange multipliers of problem (3.7). For constraint set $\mathcal{C}_2 = \{\{\hat{U}, \hat{V}\}\}$, there is only one constraint $E[Z|\hat{U}, \hat{V}]$ and on applying this, we obtain $\Lambda_{(\hat{U},\hat{V})} = -\log(E[Z|\hat{U}, \hat{V}])$ so that $\hat{Z}_B = E[Z|\hat{U}, \hat{V}]$. The minimum Bregman information solutions for all the cases are shown in the table below. Note that $\hat{Z}_B$ for the constraint set $\mathcal{C}_3$ reduces to $q(X,Y) = \frac{p(X)p(Y)p(\hat{X},\hat{Y})}{p(\hat{X})p(\hat{Y})}$ for probability distributions, which is the same as (2.3). Further, the fact that q is the minimum Bregman information solution for KL-divergence under certain constraints is equivalent to Lemma 1, which shows that is the maximum entropy distribution under those constraints.

Example 2.E (Squared Euclidean Distance) When $\phi(z) = z^2, z \in \mathbb{R}$, $\nabla\phi(z) = 2z$ and $\hat{Z}_B = -\sum_{r=1}^{s} \Lambda_r^*$ where $\Lambda^\star = \{\Lambda_r^\star\}$ are the optimal Lagrange multipliers of problem (3.7). Hence, for constraint set $\mathcal{C}_2 = \{\{\hat{U}, \hat{V}\}\}$, we obtain $\Lambda_{(\hat{U},\hat{V})} = -2E[Z|\hat{U}, \hat{V}]$ so that $\hat{Z}_B = E[Z|\hat{U}, \hat{V}]$ once again. The minimum Bregman information solutions for all the cases are shown in the table below.

3.4 A Projection Lemma

We have proposed two alternative ways, namely, (3.5) and (3.7) of quantifying the goodness of a given co-clustering (ρ, γ) with respect to a user specified constraint set $\mathcal{C}$. The following pleasantly surprising projection lemma shows that these two formulations lead to the same solution, and, henceforth, we will simply write $\hat{Z} = \hat{Z}_A = \hat{Z}_B$. The projection lemma essentially states that the minimum Bregman information solution $\hat{Z}_B$ is the Bregman projection (nearest in Bregman divergence) of Z onto the set of all approximations that satisfy the Markov property in condition A.

Lemma 2 (Projection Lemma) *For a Bregman divergence d_ϕ, any random variable $Z \in S^{m\times n}$, a specified co-clustering (ρ, γ) and a specified constraint set $\mathcal{C}$,*

$$E[d_\phi(Z, Z')] = E[d_\phi(Z, \hat{Z}_B)] + E[d_\phi(\hat{Z}_B, Z')]$$

where $Z' \in \Xi_A(\rho, \gamma, \mathcal{C})$ and $\hat{Z}_B = \hat{Z}_B(\rho, \gamma, \mathcal{C})$ as in (3.7).

Proof. By definition,

$$\begin{aligned}
&E[d_\phi(Z, Z')] \\
&\overset{(a)}{=} E[\phi(Z)] - E[\phi(Z')] - E[\langle Z - Z', \nabla\phi(Z')\rangle] \\
&= E[d_\phi(Z, \hat{Z}_B)] + E[d_\phi(\hat{Z}_B, Z')] \\
&\quad + E[\langle Z - \hat{Z}_B, \nabla\phi(\hat{Z}_B) - \nabla\phi(Z')\rangle] \\
&\overset{(b)}{=} E[d_\phi(Z, \hat{Z}_B)] + E[d_\phi(\hat{Z}_B, Z')]
\end{aligned}$$

where (a) follows from algebraic manipulation and (b) follows since $Z', \hat{Z}_B$ both satisfy Condition A and $\hat{Z}_B$ satisfies Condition B, the last term vanishes by taking conditional expectations over $\{E[Z|C],\ C \in \mathcal{C}\}$. □

Theorem 2 *For a Bregman divergence d_ϕ, any random variable $Z \in S^{m\times n}$, a specified co-clustering (ρ, γ) and a specified constraint set $\mathcal{C}$,*

$$\hat{Z}_A = \hat{Z}_B.$$

where $\hat{Z}_A$ and $\hat{Z}_B$ are given by (3.5) and (3.7) respectively.

Proof. By definition,

$$\begin{aligned}
\hat{Z}_A &= \underset{Z' \in \Xi_A(\rho,\gamma,\mathcal{C})}{\operatorname{argmin}} E[d_\phi(Z, Z')] \\
&\overset{(a)}{=} \underset{Z' \in \Xi_A(\rho,\gamma,\mathcal{C})}{\operatorname{argmin}} E[d_\phi(\hat{Z}_B, Z')] \\
&\overset{(b)}{=} \hat{Z}_B
\end{aligned}$$

where (a) follows from Lemma 2 and (b) follows since $d_\phi(\cdot,\cdot) > 0$ unless both the arguments are equal due to the strict convexity of ϕ, and the fact that $\hat{Z}_B$ satisfies condition A. □

3.5 Main Problem

The expected Bregman divergence between the given matrix Z and the minimum Bregman information solution $\hat{Z}$ provides us with a elegant way to quantify the goodness of a co-clustering. Interestingly, the following lemma shows that this expected Bregman divergence is exactly equal to the loss in Bregman information due to co-clustering, which is on the same lines as the information-theoretic co-clustering formulation as in Eqn (2.1) (originally, Lemma 2.1 in [9]).

Lemma 3 *For a Bregman divergence d_ϕ, any random variable $Z \in S^{m\times n}$, a specified co-clustering (ρ, γ) and a specified constraint set $\mathcal{C}$,*

$$E[d_\phi(Z, \hat{Z})] = I_\phi(Z) - I_\phi(\hat{Z})$$

where $\hat{Z} = \hat{Z}_A = \hat{Z}_B$ defined in (3.5) and (3.7).

Proof. By definition,

$$\begin{aligned}
&E[d_\phi(Z, \hat{Z})] \\
&= E[\phi(Z) - \phi(\hat{Z}) - \langle Z - \hat{Z}, \nabla\phi(\hat{Z})\rangle] \\
&\overset{(a)}{=} E[\phi(Z)] - E[\phi(\hat{Z})] \\
&\overset{(b)}{=} E[\phi(Z) - \phi(E[Z])] - E[\phi(\hat{Z}) - \phi(E[\hat{Z}])] \\
&\overset{(c)}{=} I_\phi(Z) - I_\phi(\hat{Z})
\end{aligned}$$

where (a) follows from the fact that $\hat{Z}$ satisfies conditions A and B so that taking conditional expectations over $\{E[Z|C], C \in \mathcal{C}\}$ makes the last term vanish, and (b) follows since $E[Z] = E[\hat{Z}]$ and (c) follows since $E[\langle Z - E[Z], \nabla\phi(E[Z])\rangle] = 0$. □

We are now ready to concretely define the generalized co-clustering problem.

Definition 1 Given k, l, a Bregman divergence d_ϕ, a data matrix $Z \in S^{m\times n}$, a set of constraints $\mathcal{C} \in \pi(\Gamma)$, and an underlying probability measure ν, we wish to find a co-clustering $(\rho^\star, \gamma^\star)$ that minimizes:

$$(\rho^\star, \gamma^\star) = \underset{(\rho,\gamma)}{\operatorname{argmin}}\ E[d_\phi(Z, \hat{Z})] = \underset{(\rho,\gamma)}{\operatorname{argmin}}\ I_\phi(Z) - I_\phi(\hat{Z}), \tag{3.8}$$

where $\hat{Z} = \hat{Z}(\rho, \gamma, \mathcal{C}) = \underset{Z' \in \Xi_B(\rho,\gamma,\mathcal{C})}{\operatorname{argmin}} I_\phi(Z')$.

The problem is NP-complete by a reduction to the `kmeans` problem. Hence, it is difficult to obtain a globally optimal solution efficiently. However, in section 4, we analyze the problem in detail, and prove that it is always possible to come up with an iterative update scheme that (a) monotonically decreases the objective function, and (b) converges to a local minimum of the problem.

Example 1.F (I-Divergence) Continuing from Example 1.C, the Bregman co-clustering objective function is given by $E[Z\log(Z/\hat{Z}) - Z + \hat{Z}] = E[Z\log(Z/\hat{Z})]$ since $E[Z] = E[\hat{Z}]$ where $\hat{Z}$ is the minimum Bregman information solution from Table 1. Note that for the constraint set $\mathcal{C}_3$ and Z based on a joint distribution $p(X, Y)$, this reduces to $D(p||q)$ where q is the joint distribution corresponding to the minimum Bregman solution indicating that (2.1) follows as a special case of (3.8).

Example 2.F (Squared Euclidean Distance) Continuing from Example 2.C, the Bregman co-clustering objective function is $E[(Z - \hat{Z})^2]$ where $\hat{Z}$ is the minimum Bregman information solution from Table 2. Note that for the constraint set $\mathcal{C}_4$, this reduces to $E[(Z - E[Z|U, \hat{V}] - E[Z|\hat{U}, V] + E[Z|\hat{U}, \hat{V}])^2]$, which is same as the objective function proposed in [6, 5].

4. A META ALGORITHM

In this section, we shall develop an alternating minimization scheme for the general Bregman co-clustering problem. Our scheme shall serve as a *meta algorithm* from which a number of special cases (both previously known and unknown) can be derived.

Throughout this section, let us suppose that the underlying measure ν, the Bregman divergence d_ϕ, the data matrix $Z \in S^{m \times n}$, number of row clusters k, number of column clusters l, and the constraint set $\mathcal{C}$ are specified and fixed. We shall focus on finding a good co-clustering for (3.8).

4.1 Intuition and Plan of Attack

We first outline the essence of our scheme.

Step 1: Start with an arbitrary row and column clustering, say, (ρ^0, γ^0). Set $t = 0$. With respect to this clustering, compute the matrix approximation $\hat{Z}^t$ by solving the minimum Bregman information problem (3.7).

Step 2: Repeat one of the following two steps till convergence:

- **Step 2A:** Hold the column clustering γ^t fixed, and find a new row co-clustering, say, ρ^{t+1}. Set $\gamma^{t+1} = \gamma^t$. With respect to co-clustering $(\rho^{t+1}, \gamma^{t+1})$, compute the matrix approximation $\hat{Z}^{t+1}$ by solving the minimum Bregman information problem. Set $t = t + 1$.
- **Step 2B:** Hold the row clustering ρ^t fixed, and find a new column co-clustering, say, γ^{t+1}. Set $\rho^{t+1} = \rho^t$. With respect to co-clustering $(\rho^{t+1}, \gamma^{t+1})$, compute the matrix approximation $\hat{Z}^{t+1}$ by solving the minimum Bregman information problem. Set $t = t + 1$.

We shall prove that this scheme converges in a finite number of steps to a local minima. Also, at any time, in Step 2, the algorithm may choose to perform either Step 2A or 2B.

4.2 A Decomposition Lemma

As is clear from the outline above, a key step in our algorithm will involve finding a solution of the minimum Bregman information problem (3.7). Besides this, we will be employing the functional form for the minimum Bregman solution $\hat{Z}$ given in Theorem 1 to obtain new matrix approximations. To be more precise, for a given $(\rho, \gamma, \mathcal{C})$, there exist a unique set of optimal Lagrange multipliers Λ^* so that Theorem 1 uniquely specifies the minimum Bregman information solution $\hat{Z}$. In general, the formula in Theorem 1 provides a unique approximation, say $\tilde{Z}$, for any set of Lagrange multipliers Λ (not necessarily optimal), and $(\rho, \gamma, \mathcal{C})$ since $\nabla\phi(\cdot)$ is a monotonic function [1, 3]. To underscore the dependence of $\tilde{Z}$ on the Lagrange multipliers, we shall use the notation $\tilde{Z} = \zeta(\rho, \gamma, \Lambda) = (\nabla\phi)^{-1}(-\sum_{r=1}^{s} \Lambda_r)$. In particular, $\hat{Z} = \hat{Z}(\rho, \gamma, \mathcal{C}) = \zeta(\rho, \gamma, \Lambda^*)$ where $\mathcal{C}$ is fixed. The basic idea in considering approximations of the form $\zeta(\rho, \gamma, \Lambda)$ is that (i) optimizing the co-clustering keeping the Lagrange multipliers fixed, and then (ii) optimizing the Lagrange multipliers, provides an efficient update scheme that does not require solving the minimum Bregman information problem anew for each possible co-clustering.

Having equipped ourselves with the above update strategy based on approximations of the form $\zeta(\rho, \gamma, \Lambda)$, we now focus on updating row clustering while keeping the column clustering fixed, and vice versa. Before we can outline concrete updates, we need an analytical tool to decompose the matrix approximation error in terms of either the rows or the columns. This *separability* makes it possible for us to efficiently obtain the best row clustering by optimizing over the individual row assignments with a fixed column clustering, and similarly for column clustering.

Lemma 4 *For a fixed co-clustering (ρ, γ) and a fixed set of (not necessarily optimal) Lagrange multipliers Λ, and $\tilde{Z} = \zeta(\rho, \gamma, \Lambda)$, we can write:*

$$\begin{aligned} E[d_\phi(Z, \tilde{Z})] &= E_U[E_{V|U}[\xi(U, \rho(U), V, \gamma(V))]] \\ &= E_V[E_{U|V}[\xi(U, \rho(U), V, \gamma(V))]], \end{aligned}$$

where $\xi(\cdot)$ is given by $\xi(U, \rho(U), V, \gamma(V)) = d_\phi(Z, \tilde{Z})$.

PROOF. By definition, $\tilde{Z} = (\nabla\phi)^{-1}(-\sum_{r=1}^{s} \Lambda_r)$. Hence,

$$\begin{aligned} E[d_\phi(Z, \tilde{Z})] &= E_{(U,V)}[d_\phi(Z, (\nabla\phi)^{-1}(-\sum_{r=1}^{s} \Lambda_r))] \\ &\overset{(a)}{=} E_{(U,V)}[\xi(U, \rho(U), V, \gamma(V))] \\ &= E_U[E_{V|U}[\xi(U, \rho(U), V, \gamma(V))]] \\ &= E_V[E_{U|V}[\xi(U, \rho(U), V, \gamma(V))]], \end{aligned}$$

where $\xi(\cdot)$ is a function determined by the Lagrange multipliers Λ, Bregman divergence d_ϕ and the original random variable Z and (a) follows since the random variables $\{C_r\}, [r]_1^s$ are subsets of $\{U, \hat{U}, V, \hat{V}\}$. $\square$

4.3 Updating Row and Column Clusters

We will now present the details of our plan in Section 4.1. First, we will demonstrate how to update row clustering (or column clustering) with respect to a fixed column clustering (or row clustering) and a fixed set of Lagrange multipliers. Then, we will find the optimal Lagrange multipliers corresponding to the minimum Bregman solution of the updated co-clustering.

Suppose we are in Step 2A outlined in Section 4.1. Updating the row clustering keeping the column clustering and the Lagrange multipliers fixed leads to a new value for the Bregman co-clustering objective function. Now making use of the separability property in Lemma 4, we can efficiently optimize the contribution of each row assignment to the overall objective function to obtain the following row cluster update step.

Lemma 5 *Let ρ^{t+1} be defined as*

$$\rho^{t+1}(u) = \underset{g:[g]_1^k}{\operatorname{argmin}}\ E_{V|u}[\xi(u, g, V, \gamma^t(V))], \quad [u]_1^m$$

and let $\tilde{Z}^t = \zeta(\rho^{t+1}, \gamma^t, \Lambda^{*t})$. Then,

$$E[d_\phi(Z, \tilde{Z}^t)] \leq E[d_\phi(Z, \hat{Z}^t)].$$

where $\hat{Z}^t = \zeta(\rho^t, \gamma^t, \Lambda^{*t})$.

Table 3: Row and column cluster updates for I-divergence.

$\mathcal{C}$	$\xi(u,g,V,\gamma(V))$	$\xi(U,\rho(U),v,h)$
$\mathcal{C}_1$	$E_{V\|u}[Z\log\left(\frac{Z}{E[Z\|g]}\right)]$	$E_{U\|v}[Z\log\left(\frac{Z}{E[Z\|h]}\right)]$
$\mathcal{C}_2$	$E_{V\|u}[Z\log\left(\frac{Z}{E[Z\|g,\hat{V}]}\right)]$	$E_{U\|v}[Z\log\left(\frac{Z}{E[Z\|\hat{U},h]}\right)]$
$\mathcal{C}_3$	$E_{V\|u}[Z\log\left(\frac{Z\times E[Z\|g]}{E[Z\|g,\hat{V}]}\right)]$	$E_{U\|v}[Z\log\left(\frac{Z\times E[Z\|h]}{E[Z\|\hat{U},h]}\right)]$
$\mathcal{C}_4$	$E_{V\|u}[Z\log\left(\frac{Z\times E[Z\|g,\hat{V}]}{E[Z\|g,V]}\right)]$	$E_{U\|v}[Z\log\left(\frac{Z\times E[Z\|\hat{U},h]}{E[Z\|U,h]}\right)]$

Table 4: Row and column cluster updates for squared Euclidean distance.

$\mathcal{C}$	$\xi(u,g,V,\gamma(V))$	$\xi(U,\rho(U),v,h)$
$\mathcal{C}_1$	$E_{V\|u}[(Z-E[Z\|g])^2]$	$E_{U\|v}[(Z-E[Z\|h])^2]$
$\mathcal{C}_2$	$E_{V\|u}[(Z-E[Z\|g,\hat{V}])^2]$	$E_{U\|v}[(Z-E[Z\|\hat{U},h])^2]$
$\mathcal{C}_3$	$E_{V\|u}[(Z-E[Z\|g,\hat{V}] + E[Z\|g])^2]$	$E_{U\|v}[(Z-E[Z\|\hat{U},h] + E[Z\|h])^2]$
$\mathcal{C}_4$	$E_{V\|u}[(Z-E[Z\|g,V] + E[Z\|g,\hat{V}])^2]$	$E_{U\|v}[(Z-E[Z\|U,h] + E[Z\|\hat{U},h])^2]$

PROOF. From Lemma 4, we have

$$\begin{aligned}
E[d_\phi(Z,\tilde{Z}^t)] &= E_U[E_{V|U}[\xi(U,\rho^{t+1}(U),V,\gamma^t(V))]] \\
&= E_U[\min_{g:[g]_1^k} E_{V|U}[\xi(U,g,V,\gamma^t(V))]] \\
&\leq E_U[E_{V|U}[\xi(U,\rho^t(U),V,\gamma(V))]] \\
&= E[d_\phi(Z,\hat{Z}^t)] \qquad \square
\end{aligned}$$

A similar argument applies to step 2B where we seek to update the column clustering keeping the row clustering fixed.

Lemma 6 *Let γ^{t+1} be defined as*

$$\gamma^{t+1}(v) = \operatorname*{argmin}_{h:[h]_1^l} E_{U|v}[\xi(U,\rho^t(U),v,h)] \quad [v]_1^n$$

and let $\tilde{Z}^t = \zeta(\rho^t,\gamma^{t+1},\Lambda^{*t})$. Then,

$$E[d_\phi(Z,\tilde{Z}^t)] \leq E[d_\phi(Z,\hat{Z}^t)].$$

where $\hat{Z}^t = \zeta(\rho^t,\gamma^t,\Lambda^{*t})$.

Applying the above Lemmas 5 and 6 for I-divergence and squared Euclidean distance, we obtain the appropriate row and column cluster updates shown in Tables 3 and 4.

Let us come back to step 2A again. So far we have only considered updating the row (or column clustering in step 2B) keeping the Lagrange multipliers fixed. After updation, the approximation $\tilde{Z}^t = \zeta(\rho^{t+1},\gamma^{t+1},\Lambda^{*t})$ is closer to the original matrix Z than the earlier minimum Bregman information solution $\hat{Z}^t$, but the Lagrange multipliers Λ^{*t} are no longer optimal and $\tilde{Z}^t$ is itself not a minimum Bregman information solution. Hence, we now optimize over the Lagrange multipliers keeping the co-clustering fixed so that the functional form $\zeta(.)$ yields the best approximation to Z. The following lemma shows that the "best" Lagrange multipliers for achieving this are the same as the optimal Lagrange multipliers of the minimum Bregman information problem.

Lemma 7 *Let $\hat{Z}^{t+1} = \zeta(\rho^{t+1},\gamma^{t+1},\Lambda^{*t+1})$ be the minimum Bregman information solution corresponding to $(\rho^{t+1},\gamma^{t+1})$ with Λ^{*t+1} being the optimal Lagrange multipliers in (3.7). Then,*

$$E[d_\phi(Z,\hat{Z}^{t+1}) \leq E[d_\phi(Z,\tilde{Z}^t)].$$

where $\tilde{Z}^t = \zeta(\rho^{t+1},\gamma^{t+1},\Lambda^{*t})$

PROOF. By definition,

$$\begin{aligned}
&E[d_\phi(Z,\hat{Z}^{t+1})] \\
&\quad= E[\phi(Z)-\phi(\hat{Z}^{t+1})-\langle Z-\hat{Z}^{t+1},\nabla\phi(\hat{Z}^{t+1})\rangle] \\
&\quad\overset{(a)}{=} E[\phi(Z)-\phi(\hat{Z}^{t+1})] \\
&\quad= E[d_\phi(Z,\tilde{Z}^t)]-E[d_\phi(\hat{Z}^{t+1},\tilde{Z}^t)]-E[\langle Z-\hat{Z}^{t+1},\nabla\phi(\tilde{Z}^t)\rangle] \\
&\quad\overset{(b)}{=} E[d_\phi(Z,\tilde{Z}^t)]-E[d_\phi(\hat{Z}^{t+1},\tilde{Z}^t)] \\
&\quad\leq E[d_\phi(Z,\tilde{Z}^t)]
\end{aligned}$$

where (a) follows since $\hat{Z}^{t+1}$ satisfies both conditions A and B so that taking conditional expectations over $E[Z|C], C\in\mathcal{C}$ makes the last term zero and (b) follows since by definition, $\nabla\phi(\tilde{Z}^t)$ is summation of terms $\Lambda_r, [r]_1^s$ and $E[\hat{Z}^{t+1}|C_r] = E[Z|C_r]$, thus making the last term vanish. $\square$

4.4 The Algorithm

Finally, we state the meta algorithm for generalized Bregman co-clustering (see Algorithm 1), that is a concrete "implementation" of our plan in Section 4.1. We now establish that our algorithm is guaranteed to achieve local optimality.

Theorem 3 *The general Bregman co-clustering algorithm (Algorithm 1) converges to a solution that is locally optimal for the Bregman co-clustering problem (3.8), i.e., the objective function cannot be improved by changing either the row clustering, the column clustering.*

PROOF. From lemmas 5, 6, and 7, it follows that updating the row clustering ρ, the column clustering γ and the Lagrange multipliers Λ one at a time decreases the objective function of the Bregman co-clustering problem. Hence, the Bregman co-clustering algorithm (Algorithm 1) which proceeds by alternately updating $\rho\to\Lambda\to\gamma\to\Lambda$ monotonically decreases the Bregman co-clustering objective function. Since the number of distinct co-clusterings is finite, the algorithm is guaranteed to converge to a locally optimal solution. Note that updation over Λ is the same as obtaining the minimum Bregman information solution. $\square$

When the Bregman divergence is I-divergence or squared Euclidean distance, the minimum Bregman information problem has a closed form analytic solution as shown in Tables 1 and 2. Hence, it is straightforward to obtain the row and column cluster update steps (Tables 3 and 4) and implement the Bregman co-clustering algorithm (Algorithm 1). The resulting algorithms involve a computational effort that is linear in the size of the data and are hence, very scalable. In general, the minimum Bregman information problem need not have a closed form solution and the update steps need to be determined using numerical computation techniques. However, since the Lagrange dual $L(\Lambda)$ in the minimum Bregman information problem (3.7) is convex in the Lagrange multipliers Λ, it is possible to obtain the optimal Lagrange multipliers using convex optimization techniques [4]. The minimum Bregman information solution and the row

Algorithm 1 Bregman Co-clustering Algorithm

Input: Matrix $Z \subseteq S^{m \times n}$, probability measure ν, Bregman divergence $d_\phi : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, num. of row clusters l, num. of column clusters k, constraint set $\mathcal{C}$.
Output: Co-clustering (ρ^*, γ^*) that (locally) optimize the objective function in (3.8).
Method:
{**Initialize** ρ, γ }
$\hat{U} \leftarrow \rho(U)$, $\hat{V} \leftarrow \gamma(V)$
repeat
{**Step A: Update Row Clusters** (ρ)}
for $u = 1$ to m **do**
$\rho(u) \leftarrow \underset{g:[g]_1^k}{\text{argmin}} \; E_{V|u}[\xi(u, g, V, \gamma(V))]$
where $\xi(U, \rho(U), V, \gamma(V)) = d_\phi(Z, \tilde{Z})$, $\tilde{Z} = \zeta(\rho, \gamma, \Lambda)$
and Λ are optimal Lagrange multipliers before updation.
end for
$\hat{U} \leftarrow \rho(U)$
{**Step B:Update Column Clusters** (γ)}
for $v = 1$ to n **do**
$\gamma(v) \leftarrow \underset{h:[h]_1^l}{\text{argmin}} \; E_{U|v}[\xi(U, \rho(U), v, h)]$
where $\xi(U, \rho(U), V, \gamma(V)) = d_\phi(Z, \tilde{Z})$, $\tilde{Z} = \zeta(\rho, \gamma, \Lambda)$
and Λ are optimal Lagrange multipliers before updation.
end for
$\hat{V} \leftarrow \gamma(V)$
until *convergence*

and column cluster update steps can then be obtained from the optimal Lagrange multipliers using Theorem 1 and Lemmas 5-7.

5. EXPERIMENTS

There are a number of experimental results in existing literature [5, 6, 9, 11] that illustrate the usefulness of particular instances of our Bregman co-clustering framework. In fact, a large class of parametric partitional clustering algorithms [3] including `kmeans` can be shown to be special cases of the proposed framework by observing that either only rows or only columns are being clustered.

In recent years, co-clustering has been successfully applied to various application domains such as text mining and analysis of microarray gene-expression data. In text mining, the information-theoretic co-clustering algorithm [9], that uses KL-divergence as the Bregman divergence, has been shown to provide superior results than "one sided" clustering algorithms that do not simultaneously cluster documents and words. Analysis of microarray gene-expression data by co-clustering of genes and experimental conditions have revealed interesting trends of gene clusters over various subsets of the experimental conditions. Since special cases of Bregman co-clustering algorithms have already been known to provide substantial improvements over other existing methods in certain domains, we do not experimentally re-evaluate the Bregman co-clustering algorithms against other methods. Instead, we present brief case studies to demonstrate four salient features of the proposed co-clustering algorithms: (a) information preserving data compression, (b) dimensionality reduction, (c) missing value prediction, and (d) learning correlations.

5.1 Information Preserving Data Compression

Bregman co-clustering provides an efficient technique to achieve compression of data matrices while preserving the specified critical statistics. When these specified statistics capture the natural structure of the data, it is possible to obtain a very accurate low parameter representation of the original data. In order to illustrate this idea, we perform co-clustering ($k = 10$, $l = 10$) on artificial 50×50 data matrices A, B, C, and D with increasing levels of complexity (produced using generative models with increasing number of parameters), with squared Euclidean distance. We present the results in Table 5 comparing co-clustering involving different sets of constraints, and different number of parameters. Clearly, for relatively simple matrices such as A and B, reasonably low parameter schemes such as $\mathcal{C}_1$ or $\mathcal{C}_2$ suffice, whereas for complicated matrices such as D, high parameter co-clustering schemes such as $\mathcal{C}_4$ seem necessary. Figure 1 shows the images of the original data matrix D, and the reconstructions obtained from each of the schemes.

Table 5: Loss in Bregman Information on matrices of increasing complexities.

Data Matrix	Co-clustering Scheme (with number of parameters)			
	$\mathcal{C}_1$ (20)	$\mathcal{C}_2$ (100)	$\mathcal{C}_3$ (200)	$\mathcal{C}_4$ (1000)
A	24.49	22.21	20.65	13.34
B	52.92	23.65	24.27	16.69
C	52.78	42.10	22.90	17.74
D	324.34	263.85	240.19	16.11

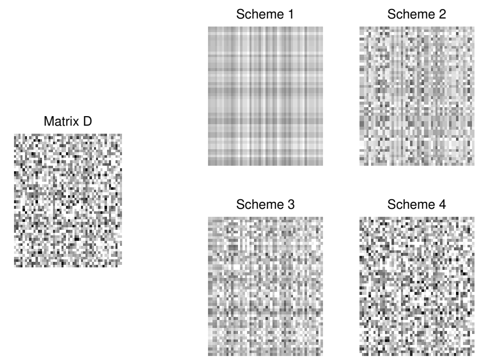

Figure 1: Matrix approximation based on various co-clustering schemes using squared Euclidean distance as the loss function.

5.2 Dimensionality Reduction

Dimensionality reduction techniques are widely used for text clustering to handle sparsity and high-dimensionality of text data. Typically, the dimensionality reduction step comes before the clustering step, and the two steps are almost independent. In practice, it is not clear which dimensionality reduction technique to use in order to get a good clustering. Co-clustering has the interesting capability of *interleaving* dimensionality reduction and clustering. This implicit dimensionality reduction results in far superior results than regular clustering techniques [9]. Due to the interleaving, stricter dimensionality reduction seems to give better results for clustering, as we show next.

Using the bag-of-words model for text, let each column be a document, and let each row be a word. Keeping the number of document clusters fixed, we present results by varying the number of word clusters. We run the experiments on 2 datasets: Classic3, a document collection from the SMART

Table 6: **Effect of Implicit Dimensionality Reduction by Co-clustering on Classic3. Each subtable is for a fixed number of (document,word) co-cluster.**

(3,20)			(3,500)			(3,2500)		
1389	1	2	1364	3	18	920	49	292
9	1455	33	5	1446	21	31	1239	404
0	4	998	29	11	994	447	172	337

Table 7: Effect of Implicit Dimensionality Reduction by Co-clustering on Different-1000. Each subtable is for a fixed number of (document,word) co-cluster.

(3,20)			(3,500)			(3,2500)		
949	15	32	435	364	146	376	368	283
31	925	80	393	454	85	251	367	363
20	60	888	172	182	769	373	265	354

project at Cornell University, and Different-1000, a subset of the benchmark 20 newsgroups dataset consisting of 3 newsgroups[4]. There are 3 classes in each dataset. Co-clustering is performed without looking at the class labels. We present confusion matrices between the cluster labels assigned by co-clustering and the true class labels, over various numbers of word clusters. The number of document clusters were fixed at 3 for all experiments reported. As we can clearly see from Table 6 (for Classic3) and Table 7 (for Different-1000), implicit dimensionality reduction by co-clustering actually gives better clusters, in the sense that the cluster labels agree more with the true class labels with fewer word clusters.

5.3 Missing Value Prediction

To illustrate missing value prediction, we consider a collaborative filtering based recommender system. The main problem in this setting is to predict the preference of a given user for a given item using the known preferences of all the users. A popular approach to handle this is by computing the Pearson correlation of each user with all other users based on the known preferences and predict the unknown rating by proportionately combining all the users' ratings. We adopt a co-clustering approach to address the same problem. The main idea is to simultaneously compute the user and item co-clusters by assigning zero measure to the missing values. As a result, the co-clustering algorithm tries to recover the original structure of the data while disregarding the missing values and the reconstructed approximate matrix can be used for prediction.

For our experimental results, we use a subset of the EachMovie dataset[5] consisting of 500 users, 200 movies and containing 25809 ratings, each rating being an integer between 0 (bad) to 5 (excellent). Of these, we use 90% ratings for co-clustering, i.e., as the training data and 10% ratings as the test data for prediction. We applied four different co-clustering algorithms ($k = 10$, $l = 10$) corresponding to constraint sets $\mathcal{C}_2$ and $\mathcal{C}_3$ with squared Euclidean (SqE) distance and I-divergence (IDiv) to the training data and used the reconstructed matrix for predicting the test ratings. We also implemented a simple collaborative filtering scheme based on Pearson's correlation. Table 8 shows the mean absolute error between the predicted ratings and the actual ratings

[4]alt.atheism, rec.sport.baseball, sci.space

[5]http://www.research.compaq.com/src/eachmovie/

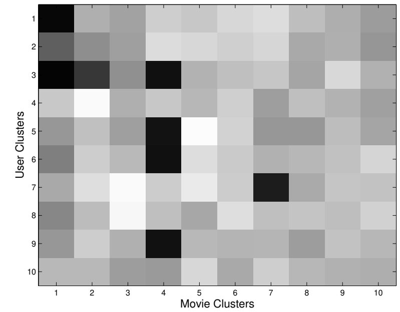

Figure 2: User-cluster Movie-cluster Correlation

for the different methods. From the table, we observe that the co-clustering achieves good results. For constraint set $\mathcal{C}_3$, the individual biases of the users (row average) and the movies (column average) are accounted for, hence resulting in a better prediction. In terms of computational effort, the co-clustering algorithms are quite efficient since the processing time is linear in the number of the known ratings.

Table 8: Mean Absolute Error for Movie Ratings

Algo.	$\mathcal{C}_2$,SqE	$\mathcal{C}_3$,SqE	$\mathcal{C}_2$,IDiv	$\mathcal{C}_3$,IDiv	Pearson
Error	0.8398	0.7639	0.8397	0.7723	1.4211

5.4 Learning Correlations

The last case-study involves discovering correlations between two sets of related entities such as genes and experimental conditions in microarray analysis, customers and products in market analysis, users and items in recommender systems, etc. We illustrate this with some anecdotal results based again on the subset of the EachMovie dataset described earlier. Figure 2 shows the average preferences of the different user clusters for the different movie clusters (dark implies higher correlation). From the figure, we observe that there are user clusters consisting of people who like lot of movies (user cluster 3) and people who like a particular kind of movies(user cluster 7). Also, there seem to be clusters of movies preferred by a lot of people (movie clusters 1 and 4) and preferred by a particular group of people (movie cluster 7). Table 9 presents a few representatives from movie clusters 1, 4 and 7. Discovering such correlations might be useful for a number of decision-making processes.

Table 9: Movie Cluster Representatives

Cluster 1	It is a Wonderful Life, Casablanca, Life is Beautiful, An Affair to Remember
Cluster 4	Usual Suspects, Manhattan Murder Mystery, Pulp Fiction, North by NorthWest
Cluster 7	Star Trek V, Blade Runner, The Terminator, A Clockwork Orange

6. RELATED WORK

Our work is primarily related to three main areas: co-clustering, matrix approximation and learning based on Bregman divergences.

Co-clustering has been a topic of much interest in the recent years because of its applications to problems such as microarray analysis [5, 6] and text mining [9]. In fact, there exist many formulations of the co-clustering problem such as the hierarchical co-clustering model [10], the biclustering model [5] that involves finding the best co-clusters one at a time, etc. In this paper, we have focussed on the partitional co-clustering formulation first introduced in [10].

Classical singular value decomposition (SVD) [14] based approaches to matrix approximation are quite often inappropriate for certain data matrices such as co-occurrence and contingency tables. Firstly, singular vectors can have negative entries. Secondly, the contributions of the component vectors in the approximation matrix are not localized. Both these issues make SVD-based decomposition difficult to interpret, which is necessary for data mining purposes. To address these and related issues, techniques involving non-negativity constraints [13] using KL-divergence as the approximation loss function [11, 13] have been proposed. However, these approaches apply to special types of matrices. A general formulation that is both interpretable and applicable to various classes of matrices seemed necessary. The proposed Bregman co-clustering formulation attempts to address this requirement.

Co-clustering involving constraints on the conditional expectations give rise to theoretically elegant models with wide range of practical applicability since key summary statistics can be naturally preserved. Several co-clustering algorithms [9, 6] that have been proposed in the recent years are derived from conditional expectation based constraints. Conditional expectation constrained co-clustering, along with its demonstrated connection to the widely used maximum entropy principle [12, 7] and conditional independence based models [11], provide a strong foundation for an unified analysis and design of unsupervised learning algorithms.

Recent research [1, 3] has shown that several results involving the KL-divergence and the squared Euclidean distance are in fact based on certain convexity properties and hence, generalize to all Bregman divergences. This intuition motivated us to consider co-clustering based on Bregman divergences. Further, the similarities between the maximum entropy and the least squares principles [8] prompted us to explore a more general minimum Bregman information principle for all Bregman divergences.

It is important to note that most clustering and co-clustering techniques based on the alternate minimization scheme can be obtained as special cases of the Bregman co-clustering algorithm. For example, the information-theoretic co-clustering [9] corresponds to the case where the constraint set is $\mathcal{C}_3$ and the Bregman divergence is KL-divergence. Similarly, the minimum sum-squared residue co-clustering algorithms [6] correspond to the cases where the constraint sets are $\mathcal{C}_1$ and $\mathcal{C}_4$ respectively and the Bregman divergence is the squared Euclidean distance. The one-sided Bregman clustering algorithms [3] are also a special case with $l = n$.

7. DISCUSSION

There are three main contributions in this paper. First, we generalized parametric co-clustering to loss functions corresponding to all Bregman divergences. The generality of the formulation makes the technique applicable to practically all types of data matrices. Second, we showed that approximation models of various complexities are possible depending on the statistics that are constrained to be preserved. Third, we proposed and extensively used the minimum Bregman information principle as a generalization of the maximum entropy principle.

For the two Bregman divergences that we focussed on, viz I-divergence and squared Euclidean distance, the proposed algorithm has linear time complexity and is hence very scalable. Scalability for other choices of Bregman divergences need to be researched. Further, like several other clustering algorithms, our algorithm only guarantees a local minima. There are well-studied techniques in the literature such as local search, deterministic annealing etc., that are capable of handling this issue and should be readily applicable to our algorithm.

8. REFERENCES

[1] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[2] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. Submitted for publication, 2003.

[3] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. In *Proc. SIAM Intl. Conf. on Data Mining*, 2004. To appear.

[4] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1998.

[5] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. 8th Intl. Conf. on Intelligent Systems for Molecular Biology (ICMB)*, pages 93–103, 2000.

[6] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum sum squared residue co-clustering of gene expression data. In *Proc. 4th SIAM International Conference on Data Mining (SDM)*, 2004. To appear.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[8] I. Csiszar. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19:2032–2066, 1991.

[9] I. Dhillon, S. Mallela, and D. Modha. Information-theoretic co-clustering. In *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 89–98, 2003.

[10] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[11] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report ICSI TR-98-042, International Computer Science Institute (ICSI), Berkeley, 1998.

[12] E. T. Jaynes. Information theory and statistical mechanics. *Physical Reviews*, 106:620–630, 1957.

[13] D. L. Lee and S. Seung. Algorithms for non-negative matrix factorization. In *Proc. 14th Ann. Conf. on Neural Information Processing Systems (NIPS)*, 2001. 556-562.

[14] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proc. 16th Ann. ACM Symposium on Principles of Distributed Computing (PODC)*, 1998.