Biclustering with Alternating K-Means

Nicolas Fraiman
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3260, USA

Zichao Li
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3260, USA

Abstract

Biclustering is the task of simultaneously clustering the rows and columns of the data matrix into different subgroups such that the rows and columns within a subgroup exhibit similar patterns. In this paper, we consider the case of producing exclusive row and column biclusters. We provide a new formulation of the biclustering problem based on the idea of minimizing the empirical clustering risk. We develop and prove a consistency result with respect to the empirical clustering risk. Since the optimization problem is combinatorial in nature, finding the global minimum is computationally intractable. In light of this fact, we propose a simple and novel algorithm that finds a local minimum by alternating the use of an adapted version of the $k$-means clustering algorithm between columns and rows. We evaluate and compare the performance of our algorithm to other related biclustering methods on both simulated data and real-world gene expression data sets. The results demonstrate that our algorithm is able to detect meaningful structures in the data and outperform other competing biclustering methods in various settings and situations.

Keywords: biclustering, co-clustering, $k$-means clustering, empirical risk minimization, gene expression analysis

1. Introduction

In many fields of application, the data can be represented by a matrix, and people are interested in the task of simultaneously clustering the rows and columns of the data matrix into different subgroups such that the rows and columns within a subgroup exhibit similar patterns. This general task has been studied in many different application domains. For example, in gene expression analysis, people seek to identify subgroups of genes that have similar expression levels within corresponding subgroups of conditions (Cheng and Church, 2000). In text mining, people attempt to recognize subgroups of documents that have similar properties with respect to corresponding subgroups of words (Dhillon, 2001). In collaborative filtering, people wish to detect subgroups of customers with similar preferences toward corresponding subgroups of products (Hofmann and Puzicha, 1999). The most common name of the task is biclustering (Cheng and Church, 2000; Tanay et al., 2002; Yang et al., 2003; Sheng et al., 2003; Kluger et al., 2003; Prelić et al., 2006), although it is also known by other names such as co-clustering (Dhillon, 2001; Hanisch et al., 2002; Dhillon et al., 2003; Cho et al., 2004; Banerjee et al., 2007; Seldin and Tishby, 2010).
Fraiman and Li

subspace clustering (Agrawal et al., 1998; Parsons et al., 2004; Vidal, 2011; Elhamifar and Vidal, 2013), and direct clustering (Hartigan, 1972).

Over the years, a large number of biclustering methods have been proposed. Comprehensive reviews of different biclustering methods can be found in Madeira and Oliveira (2004) and Tanay et al. (2005). The biclustering methods could be classified into different groups based on the structure of the produced biclusters. Figure 1 shows three types of bicluster structures that could be obtained after appropriate row and column reordering:

1. In Figure 1a, the biclusters are arbitrarily positioned and can overlap with each other. The majority of the biclustering methods produce this type of biclusters, including Cheng and Church (Cheng and Church, 2000), CTWC (Getz et al., 2000), ISA (Bergmann et al., 2003), SAMBA (Tanay et al., 2004), plaid models (Lazzeroni and Owen, 2002), OPSM (Ben-Dor et al. 2002), xMOTIFs (Murali and Kasif, 2002), and others (Li et al., 2009; Shabalin et al., 2009; Hochreiter et al., 2010).

2. In Figure 1b, the biclusters are non-overlapping and follow a checkerboard structure. The biclustering methods that produce this type of biclusters include spectral biclustering (Kluger et al., 2003), SSVD (Lee et al., 2010), sparse biclustering (Tan and Witten, 2014), convex biclustering (Chi et al., 2017), profile likelihood biclustering (Flynn et al., 2020), and others (Dhillon, 2001; Cho et al., 2004; Chen et al., 2013).

3. In Figure 1c, the biclusters are rectangular diagonal blocks in the data matrix. In this case, there exist $k$ mutually exclusive and exhaustive clusters of rows, and $k$ corresponding mutually exclusive and exhaustive clusters of columns. Our method, named alternating $k$-means biclustering, produces this type of biclusters.

Some biclustering methods (Segal et al., 2001; Tang et al., 2001; Wang et al., 2002; Gu and Liu, 2008; Sill et al., 2011) produce other types of bicluster structures. A more detailed discussion is provided in Madeira and Oliveira (2004). In general, methods that produce overlapping biclusters are more complex and flexible, whereas methods that produce non-overlapping biclusters are easier to interpret and visualize.
The key difference between clustering and biclustering is that clustering is based on global patterns whereas biclustering is based on local patterns. More specifically, when performing clustering on the rows of the data matrix, all the columns are taken into consideration. In contrast, when performing biclustering on the data matrix, the rows are clustered into different groups based on different subsets of columns. This characteristic of biclustering inspired us to develop a “local” version of \( k \)-means clustering: instead of performing \( k \)-means clustering on the rows using all the columns, we could consider only using different subsets of columns. In other words, the cluster centers could be defined using different subsets of columns instead of all the columns. Starting from this simple idea, we adapt the formulation and algorithm of \( k \)-means clustering to a biclustering method by making several important modifications, and the details of our formulation and algorithm are given in Section 2 and Section 5, respectively. Notably, our alternating \( k \)-means biclustering algorithm is conceptually as simple as \( k \)-means clustering, and it has the extra advantage of being able to discover local patterns as opposed to global patterns. These two characteristics make our algorithm an ideal candidate to serve as a baseline for biclustering problems even when our bicluster structure might not be flexible enough.

Our main contributions can be summarized as follows:

1. We provide a new formulation of the biclustering problem based on the idea of minimizing the empirical clustering risk. The formulation is adapted from the \( k \)-means clustering problem, with two important changes with respect to the definitions of cluster centers and norm. We further develop and prove a consistency result with respect to the empirical clustering risk, which is generally quite rare for biclustering methods.

2. Since minimizing the empirical clustering risk is a combinatorial optimization problem, finding the global minimum is computationally intractable. In light of this fact, we propose a simple and novel algorithm that finds a local minimum by alternately applying an adapted version of the \( k \)-means clustering algorithm between columns and rows. The simplicity of our algorithm makes it easy to understand, implement, and interpret. The R package \texttt{akmbiclust}, available on CRAN, implements our alternating \( k \)-means biclustering algorithm.

3. We empirically evaluate and compare the performance of our method to other related biclustering methods on both simulated data and real-world gene expression data sets. The empirical results have demonstrated that our method is able to detect meaningful structures in the data and outperform other competing biclustering methods in various settings and situations.

The rest of this paper is organized as follows. In Section 2, we formulate the task of biclustering as an optimization problem and present a consistency result. In Section 3, we provide a rigorous proof of the consistency result. In Section 4, we describe a probabilistic interpretation of the optimization problem. In Section 5, we present a simple algorithm that finds the local optimum by alternating the use of \( k \)-means clustering between columns and rows. In Section 6, we propose extending our method by adding penalization terms. In Section 7, we evaluate and compare the algorithm’s performance on simulated data to other related biclustering algorithms. In Section 8, we apply the algorithm to three cancer gene
expression data sets, demonstrating its advantage over other related biclustering algorithms in terms of sample misclassification rate. In Section 9, we conclude with a discussion.

2. Problem Formulation and Consistency Result

Suppose we have a \( n \times m \) matrix \( X \) representing \( n \) data points \( X_1, \ldots, X_n \in \mathbb{R}^m \). A typical example is the gene expression matrix, with rows corresponding to genes and columns corresponding to conditions. We formulate the task of biclustering on \( X \) as partitioning the \( n \) rows and \( m \) columns into \( k \) groups to get \( k \) biclusters. More specifically, let \( J = \{1, \ldots, n\} \) be the set of row indices, then \( J \) could be partitioned into \( k \) disjoint nonempty sets \( J_1, \ldots, J_k \), where \( J_1 \cup \cdots \cup J_k = J \). Similarly, let \( I = \{1, \ldots, m\} \) be the set of column indices, then \( I \) could also be partitioned into \( k \) disjoint nonempty sets \( I_1, \ldots, I_k \), where \( I_1 \cup \cdots \cup I_k = I \). The \( k \) groups of row indices \( J_1, \ldots, J_k \) and column indices \( I_1, \ldots, I_k \) could be viewed as \( k \) biclusters. Note that under this definition of biclustering every row and column in the matrix \( X \) belongs to one and only one bicluster. In other words, the rows and columns in the biclusters are exhaustive and exclusive.

For any \( X = (x_1, \ldots, x_m) \in \mathbb{R}^m \), let \( X(I_j) = (x_i)_{i \in I_j} \). For example, let \( X = (1, 3, 4, 7) \), and \( I_1 = \{1, 3\}, I_2 = \{2, 4\} \). Then \( X(I_1) = (1, 4), X(I_2) = (3, 7) \). The space of \( X(I_j) \) is defined as \( \mathbb{R}^{I_j} \). We define a special norm on \( \mathbb{R}^{I_j} \), called dimensionality-normalized norm. For any \( X \in \mathbb{R}^{I_j} \), let \( l_j = |I_j| \) denote the cardinality of the index set \( I_j \), then it is also the dimension of the space \( \mathbb{R}^{I_j} \), and the dimensionality-normalized norm of \( X \) is defined as

\[
||X||_{dn} = \sqrt{\frac{\sum_{i \in I_j} x_i^2}{l_j}}.
\]

The name “dimensionality-normalized norm” comes from the following simple relationship between the dimensionality-normalized norm and the Euclidean norm:

\[
||X||^2_{dn} = \frac{||X||^2}{l_j}.
\]

Our method seeks to find the \( k \) groups of column indices \( I_j, 1 \leq j \leq k \) and the \( k \) cluster centers \( c_j \in \mathbb{R}^{I_j}, 1 \leq j \leq k \) such that the following objective function is minimized:

\[
\sum_{i=1}^{n} \min_{1 \leq j \leq k} ||X_i(I_j) - c_j||^2_{dn}.
\]

The corresponding \( k \) groups of row indices \( J_t, 1 \leq t \leq k \) can be obtained by selecting all the rows that are “closest” to cluster center \( c_t \):

\[
J_t = \{i : \text{arg min}_{1 \leq j \leq k} ||X_i(I_j) - c_j||^2_{dn} = t\}, 1 \leq t \leq k.
\]

Note that here “closest” is measured by the distance function induced by the dimensionality-normalized norm:

\[
dist(X_i(I_j), c_j) = ||X_i(I_j) - c_j||_{dn}, 1 \leq j \leq k, 1 \leq i \leq n.
\]
Our biclustering method can be viewed as a more complicated version of the traditional $k$-means clustering, which only seeks to find the $k$ cluster centers $c_j \in \mathbb{R}^m, 1 \leq j \leq k$ such that the following objective function is minimized:

$$
\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i - c_j\|_2^2.
$$

However, it is important to note that there are two key differences:

1. The $k$ cluster centers $c_1, \ldots, c_k$ in our objective function are not vectors in $\mathbb{R}^m$. Instead, $c_j \in \mathbb{R}^{I_j}$ for $1 \leq j \leq k$, and the $k$ groups of column indices $I_1, \ldots, I_k$ also are parameters that we need to optimize over. In fact, finding the best column partition $I_1, \ldots, I_k$ is combinatorial in nature, which makes the optimization problem computationally intractable.

2. The norm in our objective function is not the Euclidean norm. Instead, it is the dimensionality-normalized norm.

Suppose the data is a sequence of independent random observations $X_1, \ldots, X_n \in \mathbb{R}^m$ with the same distribution as a generic random variable $X$ with distribution $\mu$. We minimize the empirical clustering risk

$$
W(I, c, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i(I_j) - c_j\|_{dn}^2
$$

over all possible choices of column partitions $I = \{I_j\}_{1 \leq j \leq k}$ and cluster centers $c = \{c_j\}_{1 \leq j \leq k}$. Here, $\mu_n$ is the empirical distribution of the data.

The performance of a clustering scheme given by the column partition $I$ and cluster centers $c$ is measured by the clustering risk

$$
W(I, c, \mu) = \int \min_{1 \leq j \leq k} \|x(I_j) - c_j\|_{dn}^2 d\mu(x).
$$

The optimal clustering risk is defined as

$$
W^*(\mu) = \inf_I \inf_c W(I, c, \mu).
$$

Let $\delta_n \geq 0$. A column partition $I_n$ and cluster centers $c_n$ as a whole is a $\delta_n$-minimizer of the empirical clustering risk if

$$
W(I_n, c_n, \mu_n) \leq W^*(\mu_n) + \delta_n,
$$

where $W^*(\mu_n) = \inf_I \inf_c W(I, c, \mu_n)$. When $\delta_n = 0$, $I_n$ and $c_n$ as a whole is called an empirical risk minimizer. Since $\mu_n$ is supported on at most $n$ points, the existence of an empirical risk minimizer is guaranteed.

The key theoretical result of this paper is the following consistency theorem, which states that the clustering risk of a $\delta_n$-minimizer of the empirical clustering risk converges to the optimal risk as long as $\lim_{n \to \infty} \delta_n = 0$. 

Theorem 1 Assume that $\mathbb{E}|X|^2 < \infty$. Let $I_n$ and $c_n$ be a $\delta_n$-minimizer of the empirical clustering risk. If $\lim_{n \to \infty} \delta_n = 0$, then

1. $\lim_{n \to \infty} W(I_n, c_n, \mu) = W^*(\mu) \text{ a.s.},$ and
2. $\lim_{n \to \infty} \mathbb{E}W(I_n, c_n, \mu) = W^*(\mu)$.

It is important to point out that we assume a minimizer of the empirical clustering risk can be found. However, finding the global minimum of the empirical clustering risk is a computationally intractable problem due to its combinatorial nature. In light of this fact, we present a simple algorithm in Section 5 that finds a local minimum based on the idea of alternating the use of $k$-means clustering between columns and rows.

3. Proof of the Consistency Result

In this section, we prove the consistency result of Theorem 1. We mainly follow the steps in Biau et al. (2008). To prove Theorem 1, we need to prove three lemmas. Recall that the $L_2$ Wasserstein distance between two probability measures $\mu_1$ and $\mu_2$ on $\mathbb{R}^m$, with finite second moment, is defined as

$$\gamma(\mu_1, \mu_2) = \inf_{X \sim \mu_1, Y \sim \mu_2} (\mathbb{E}|X - Y|^2)^{1/2},$$

where the infimum is taken over all joint distributions of two random variables $X$ and $Y$ such that $X$ has distribution $\mu_1$ and $Y$ has distribution $\mu_2$. It has been shown in Linder (2002) that $\gamma$ is a metric on the space of probability distributions on $\mathbb{R}^m$ with finite second moment, and that the infimum is a minimum and can be achieved.

Lemma 2 shows that if two distributions $\mu_1$ and $\mu_2$ are close in terms of $L_2$ Wasserstein distance, then their clustering risks are also similar.

Lemma 2 For any column partition $I$ and cluster centers $c$,

$$\left| W(I, c, \mu_1)^{1/2} - W(I, c, \mu_2)^{1/2} \right| \leq \gamma(\mu_1, \mu_2).$$

Proof Let $X \sim \mu_1$ and $Y \sim \mu_2$ achieve the infimum defining $\gamma(\mu_1, \mu_2)$. Then

$$W(I, c, \mu_1)^{1/2} = \left[ \int \min_{1 \leq j \leq k} ||x(I_j) - c_j||^2_{d_n} d\mu_1(x) \right]^{1/2}$$

$$= \left[ \mathbb{E} \min_{1 \leq j \leq k} ||X(I_j) - c_j||^2_{d_n} \right]^{1/2}$$

$$\leq \left[ \mathbb{E} \min_{1 \leq j \leq k} (||X(I_j) - Y(I_j)||^2_{d_n} + ||Y(I_j) - c_j||^2_{d_n}) \right]^{1/2}$$

$$\leq \left[ \mathbb{E} \min_{1 \leq j \leq k} (||X(I_j) - Y(I_j)||^2 + ||Y(I_j) - c_j||^2_{d_n}) \right]^{1/2}$$

$$\leq \left[ \mathbb{E}||X - Y||^2_{d_n} \right]^{1/2} + \left[ \mathbb{E} \min_{1 \leq j \leq k} ||Y(I_j) - c_j||^2_{d_n} \right]^{1/2}$$

$$= \gamma(\mu_1, \mu_2) + W(I, c, \mu_2)^{1/2},$$
which implies that $W(I, c, \mu_1)^{1/2} - W(I, c, \mu_2)^{1/2} \leq \gamma(\mu_1, \mu_2)$. The other direction can be proved similarly.

Lemma 3 relates the clustering risk $W(I_n, c_n, \mu)$ of a $\delta_n$-minimizer of the empirical clustering risk to the optimal risk $\inf_I \inf_c W(I, c, \mu)$ in terms of the $L_2$ Wasserstein distance between $\mu$ and $\mu_n$.

**Lemma 3** Let $I_n$ and $c_n$ be a $\delta_n$-minimizer of the empirical clustering risk. Then

$$W(I_n, c_n, \mu)^{1/2} - \left[\inf_I \inf_c W(I, c, \mu)\right]^{1/2} \leq 2\gamma(\mu, \mu_n) + \sqrt{\delta_n}.$$ 

**Proof** Let $\varepsilon > 0$ be arbitrary, and let $I^*$ and $c^*$ be any element satisfying

$$\inf_I \inf_c W(I, c, \mu) \leq W(I^*, c^*, \mu) < \inf_I \inf_c W(I, c, \mu) + \varepsilon.$$ 

For any $t \in \mathbb{R}$, we set $(t)_+ = \max(t, 0)$. Then

$$W(I_n, c_n, \mu)^{1/2} - \left[\inf_I \inf_c W(I, c, \mu)\right]^{1/2} \leq W(I_n, c_n, \mu)^{1/2} - W(I^*, c^*, \mu)^{1/2} + \varepsilon.$$ 

where the last inequality follows from Lemma 2.

Lemma 4 states that the $L_2$ Wasserstein distance between $\mu$ and $\mu_n$ converges to 0.

**Lemma 4**

1. $\lim_{n \to \infty} \gamma(\mu, \mu_n) = 0$ a.s., and

2. $\lim_{n \to \infty} \mathbb{E}\gamma^2(\mu, \mu_n) = 0$.

**Proof**

1. A well known result is that the empirical measure $\mu_n$ converges to $\mu$ almost surely. By Skorokhod’s representation theorem, there exist $Y_n \sim \mu_n$ and $Y \sim \mu$ jointly distributed such that $Y_n \to Y$ a.s. By the triangle inequality, we have $2||Y_n||_2^2 + 2||Y||_2^2 - ||Y_n - Y||_2^2 \geq ||Y_n||_2^2 + ||Y||_2^2 - 2||Y_n||_2||Y||_2 \geq 0$. Hence Fatou’s lemma implies

$$\lim_{n \to \infty} \mathbb{E}\left[2||Y_n||_2^2 + 2||Y||_2^2 - ||Y_n - Y||_2^2\right] \geq \mathbb{E}\left[2||Y_n||_2^2 + 2||Y||_2^2 - ||Y_n - Y||_2^2\right] = 4\mathbb{E}||Y||_2^2.$$ 

Since $\lim_{n \to \infty} \mathbb{E}||Y_n||_2^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} ||X_i||_2^2 = \mathbb{E}||X||_2^2 = \mathbb{E}||Y||_2^2$, we must have $\lim_{n \to \infty} \mathbb{E}||Y_n - Y||_2^2 = 0$, which implies that $\lim_{n \to \infty} \gamma(\mu, \mu_n) = 0$ almost surely.
2. Let $\mathcal{M}(\mu, \mu_n)$ denote the set of all laws on $\mathbb{R}^m \times \mathbb{R}^m$ with marginals $\mu$ and $\mu_n$. By definition, the squared $L_2$ Wasserstein distance between $\mu$ and $\mu_n$ can be written as

$$\gamma^2(\mu, \mu_n) = \inf_{\nu \in \mathcal{M}(\mu, \mu_n)} \int ||x - y||_2^2 d\nu(x, y).$$

Let $C$ be an arbitrary nonnegative constant, and let $\mathcal{A}$ be the subset of $\mathbb{R}^m \times \mathbb{R}^m$ defined by

$$\mathcal{A} = \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^m : \max(||x||_2, ||y||_2) \leq C\}.$$

For any $\nu \in \mathcal{M}(\mu, \mu_n)$, we have

$$\int ||x - y||_2^2 d\nu(x, y) = \int_A ||x - y||_2^2 d\nu(x, y) + \int_{A^c} ||x - y||_2^2 d\nu(x, y)$$

$$\leq \int_A ||x - y||_2^2 d\nu(x, y) + 2 \int_{A^c} ||x||_2^2 d\nu(x, y) + 2 \int_{A^c} ||y||_2^2 d\nu(x, y)$$

$$\leq \int_A ||x - y||_2^2 d\nu(x, y)$$

$$+ 2 \int ||x||_2^2 I_{\{||x||_2 > C\}} d\mu(x) + 2 \int ||y||_2^2 I_{\{||y||_2 \leq C\}} d\nu(x, y)$$

$$\leq \int_A ||x - y||_2^2 d\nu(x, y)$$

$$+ 2 \int ||x||_2^2 I_{\{||x||_2 > C\}} d\mu(x) + 2 = \mu_n \{||y||_2 > C\}$$

$$\leq \int_A ||x - y||_2^2 d\nu(x, y)$$

$$+ 2 \int ||x||_2^2 I_{\{||x||_2 > C\}} d\mu(x) + 2 \int ||y||_2^2 I_{\{||y||_2 \leq C\}} d\nu(x, y)$$

$$\leq \int_A ||x - y||_2^2 d\nu(x, y)$$

$$+ 2 \int ||x||_2^2 I_{\{||x||_2 > C\}} d\mu(x) + 2 \int ||y||_2^2 I_{\{||y||_2 \leq C\}} d\nu(x, y)$$

where the last inequality follows from the Markov’s inequality. Taking the infimum over $\mathcal{M}(\mu, \mu_n)$ on both sides and taking expectations with respect to the $X_i$’s, we have

$$\mathbb{E}\gamma^2(\mu, \mu_n) \leq \mathbb{E} \left[ \inf_{\nu \in \mathcal{M}(\mu, \mu_n)} \int_A ||x - y||_2^2 d\nu(x, y) \right] + 8 \int ||x||_2^2 I_{\{||x||_2 > C\}} d\mu(x).$$

For a fixed $C \geq 0$, the first term of the right-hand side goes to 0 as $n \to \infty$ according to part 1 of this lemma and the Lebesgue dominated convergence theorem. Since $\int ||x||_2^2 d\mu(x) < \infty$, the second term of the right-hand side goes to 0 as $C \to \infty$, and this concludes the proof of part 2 of this lemma.

The two statements of Theorem 1 are immediate consequences of Lemma 3 and 4.
4. A Probabilistic Interpretation of the Optimization Problem

In this section, we provide a probabilistic interpretation of the optimization problem, which also serves as the motivation behind the definition of the dimensionality-normalized norm. Suppose every row $X_i$ in the matrix $X$ is generated independently through the following process:

1. Select a nonempty subset of all the columns $I_j \subset I$, and the entries in those columns follow a multivariate normal distribution with mean vector $c_j$ and covariance matrix $\sigma^2 I$ (here $I$ denotes the identity matrix):
   \[ X_i(I_j) \sim \mathcal{N}(c_j, \sigma^2 I). \]

2. The entries in the other columns are considered as noise and do not affect the likelihood of $X_i$.

The log-likelihood of $X_i$ has the following property:

\[ \log L(I_j, c_j | X_i) \propto -\frac{1}{2\sigma^2} \|X_i(I_j) - c_j\|^2 - \frac{l_j}{2} \log(2\pi \sigma^2), \]

where $l_j = |I_j|$ is the cardinality of the index set $I_j$ and also the dimensionality of the vector $X_i(I_j)$.

Naturally, the next step is to maximize the log-likelihood of $X_i$ over $I_j$ and $c_j$. However, there is one issue: different $I_j$ might have different cardinality $l_j$. This means that $X_i(I_j)$ might have different dimensionality, and directly comparing the log-likelihood of vectors of different dimensionality is problematic: when $\log(2\pi \sigma^2) > 0$, increasing the dimensionality $l_j$ would monotonically decrease the log-likelihood.

One simple solution is to maximize the dimensionality-normalized log-likelihood of $X_i$:

\[ \frac{\log L(I_j, c_j | X_i)}{l_j} \propto -\frac{\|X_i(I_j) - c_j\|^2}{l_j}, \]

which is equivalent to minimizing

\[ \frac{\|X_i(I_j) - c_j\|^2}{l_j} = \|X_i(I_j) - c_j\|^2_{dn}. \]

Therefore maximizing the joint dimensionality-normalized log-likelihood of all the rows in the matrix $X$ is equivalent to minimizing the empirical clustering risk

\[ \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i(I_j) - c_j\|^2_{dn}. \]

This equivalence establishes a connection between the optimization perspective and the probabilistic perspective of the biclustering problem.
5. Algorithm

In this section, we present a simple and novel algorithm that finds a local minimum of the empirical clustering risk. The idea is to alternate the use of an adapted version of the $k$-means clustering algorithm between columns and rows. Similar to the widely used Lloyd’s algorithm in $k$-means clustering, our algorithm is also based on heuristics and does not guarantee to achieve global optimum. In each individual run, our alternating $k$-means biclustering algorithm works as described in Algorithm 1.

**Algorithm 1** Alternating $k$-means biclustering

1. Start by performing $k$-means clustering separately on the rows and columns of the input matrix $X$ to obtain the initial partitions of the $n$ rows $J_1, \ldots, J_k$ and $m$ columns $I_1, \ldots, I_k$. Calculate and record the loss.

2. With a fixed $I_1, \ldots, I_k$, the optimal cluster centers $c_1, \ldots, c_k$ could be found in the following way:

   (a) (Update step) Given row partitions $J_1, \ldots, J_k$, update the cluster centers $c_1, \ldots, c_k$ by the following equation:

   $$c_j = \frac{1}{|J_j|} \sum_{i \in J_j} X_i(I_j), 1 \leq j \leq k.$$  

   (b) (Assignment step) Given cluster centers $c_1, \ldots, c_k$, update the row partitions $J_1, \ldots, J_k$ by assigning every row to the cluster center with the smallest distance (induced by the dimensionality-normalized norm), and all the rows that are closest to $c_j$ form $J_j, 1 \leq j \leq k$.

   Alternate between (a) and (b) until convergence, and obtain a partition of the $n$ rows $J_1, \ldots, J_k$.

3. Transpose the matrix $X$, and $J_1, \ldots, J_k$ becomes a partition of the $n$ columns. Again, alternate between (a) and (b) until convergence, and obtain a partition the $m$ rows $I_1, \ldots, I_k$. Transpose the matrix $X$ back, and $I_1, \ldots, I_k$ becomes a partition of the $m$ columns.

4. Alternate between step 2 and step 3 until convergence. Calculate and record the loss.

5. Compare the losses at the end of step 1 and step 4. Output the $k$ groups of row indices $J_1, \ldots, J_k$ and column indices $I_1, \ldots, I_k$ associated with the minimum loss.

Noticeably, the subroutine that alternates between (a) and (b) is quite similar to the widely used Lloyd’s algorithm in $k$-means clustering, which is the reason why our algorithm is called “alternating $k$-means biclustering”. However, we note that there are two important differences:

1. The cluster centers $c_1, \ldots, c_k$ are not vectors in $\mathbb{R}^m$. Instead, $c_j \in \mathbb{R}^{I_j}$ for $1 \leq j \leq k$. 


2. When calculating the distance between a row $X_i$ and a cluster center $c_j$, the distance function is not induced by the Euclidean norm. Instead, it is induced by the dimensionality-normalized norm.

It is recommended to run our algorithm multiple times and choose the result with the minimum loss, each time starting with a different initialization by randomly permuting the rows and columns of the input matrix $X$. The reason is twofold:

1. First, our algorithm does not guarantee global optimum and depends on the partitions obtained from the initial separate $k$-means clustering, so running our algorithm multiple times increases the probability of finding a smaller local minimum.

2. Second, just like $k$-means clustering algorithm, our algorithm also might encounter empty cluster problem. More specifically, if in any assignment step any group of row indices $J_j, 1 \leq j \leq k$ becomes empty, then the algorithm cannot proceed and need to restart. The probability of having empty clusters is small when $k$ is much smaller than $\min(n, m)$, but might become larger when $k$ approaches $\min(n, m)$.

An implementation of our algorithm is provided in the R package akmbiclust, available on CRAN.

6. Penalization

In this section, we consider extending our method by adding penalization terms to the loss function. So far, the loss function that we minimize is the empirical clustering risk:

$$
\frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||X_i(I_j) - c_j||^2_{d_n}.
$$

Intuitively, minimizing this loss encourages all the rows in the bicluster $j$ to have similar entries in the columns $I_j$ across different rows. However, it does not differentiate between how large or small those entries are. Therefore, it might be beneficial to place some form of penalization on those entries to encourage "good" biclustering results. To this end, we consider the following penalization method:

Let $||X||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2}$ denote the Frobenius norm of a matrix $X \in \mathbb{R}^{n \times m}$. Let the index of the $k$ biclusters be from 0 to $k-1$. For every bicluster $j$ where $0 \leq j \leq k-1$, let $X^{(j)}$ denote the submatrix of $X$ consisting only of rows and columns that belong to bicluster $j$. The following penalization term is added to every bicluster $j$ where $1 \leq j \leq k-1$:

$$
\lambda \cdot \frac{||X||^2_F}{||X^{(j)}||^2_F + 1}.
$$

Note that bicluster 0 is the special bicluster that does not have the above penalization term. The parameter $\lambda$ is a tuning parameter.

The motivation of the penalization method comes from the following two observations:

1. In general, the entries in the biclusters should represent signals, which means that they should not be close to zero. Therefore the Frobenius norm of the submatrix $||X^{(j)}||_F$ induced by the bicluster $j$ should be large.
2. However, not all rows and columns should be classified into one of $k$ biclusters representing signals. Some rows or columns might just consist of random noise, and it is reasonable to include a special bicluster that represents random noise. Adding the penalization term $\lambda \cdot \frac{\|X\|^2_{F}}{\|X\|^2_{F} + 1}$ to every bicluster creates a potential problem: when you are supposed to have a bicluster with entries close to zero, the penalization term might become excessively large. Thus we choose to not apply the penalization term to bicluster 0, which is the special bicluster representing noise.

The penalized loss function can be written as:

$$\frac{1}{n} \sum_{i=1}^{n} \min_{0 \leq j \leq k-1} \|X_i(I_j) - c_j\|^2_{dn} + \lambda \sum_{j=1}^{k-1} \frac{\|X\|^2_{F}}{\|X\|^2_{F} + 1}.$$  

When $\lambda = 0$, the penalized loss function reduces to the empirical clustering risk.

It is important to point out that the loss function does not affect the alternating process between step 2 and step 3 in Algorithm 1. However, the loss function does affect which specific biclustering result is chosen among different biclustering results produced by the algorithm: typically the algorithm is run with many random initializations, and even in each individual run, the algorithm needs to compare the two losses of two biclustering results: one at the end of step 1 corresponding to the partitions obtained by separate $k$-means, one at the end of step 4 when the algorithm finishes performing alternating $k$-means. Among all the different biclustering results, the biclustering result with the minimum loss is chosen as the final output.

7. Simulation Studies

In this section, we evaluate and compare the performance of the following four biclustering methods on simulated data with different settings:

1. Alternating $k$-means biclustering (AKM): This is the method presented in this paper. We use the penalized loss function in Section 6, with three different $\lambda$ values: 0, 0.1, and 1.

2. Separate $k$-means clustering (KM): This method simply performs $k$-means clustering separately on the rows and columns.

3. Profile likelihood biclustering (PL) (Flynn et al., 2020): This method is based on profile likelihood and has associated consistency guarantees. We implement the method using the R package biclustpl. The distribution family is selected as Gaussian, which is the true distribution of the simulated data.

4. Sparse biclustering (SBC) (Tan and Witten, 2014): This method assumes the entries are normally distributed with a bicluster-specific mean and a common variance, and maximizes the $L_1$-penalized log-likelihood to obtain sparse biclusters. We implement the method using the R package sparseBC. The input matrix is always mean-centered before applying the method, and the tuning parameter $\lambda$ is selected by choosing the $\lambda$ with the smallest BIC over a grid of $\lambda$ values, both of which are suggested in their paper.
All methods are run with 100 random initializations.

Among numerous existing biclustering methods, the above four methods are selected to evaluate and compare their performance in the simulation studies because they satisfy the following two requirements:

1. Every row should be classified into one and only one row cluster. In addition, every column should also be classified into one and only one column cluster. This means that the biclustering methods should produce non-overlapping biclusters with checkerboard structure (Figure 1b) or exclusive row and column biclusters (Figure 1c).

2. In addition, the biclustering methods should also allow explicitly specifying the number of clusters that the rows and columns are classified into. A few biclustering methods such as spectral biclustering (Kluger et al., 2003), SSVD (Lee et al., 2010), and convex biclustering (Chi et al., 2017) satisfy the first requirement but do not satisfy this requirement.

In our simulations, the evaluation metric is the misclassification rate, which is defined as:

\[
\text{misclassification rate} = \frac{\text{number of entries classified into the wrong row or column cluster}}{\text{total number of entries in the input matrix } X}. 
\]

Smaller misclassification rate indicates better performance, and a perfect biclustering result would have a misclassification rate of 0.

We generate simulated data in three different settings. In all three settings, the input matrix \( X \) is generated using a \( 2 \times 2 \) block model, and for all the methods we set the number of clusters that the rows and columns are classified into to be 2. The number of rows is set to be \( n = 400 \), and the number of columns is set to be \( m = a \cdot n \) where \( a \in \{0.5, 1.0, 2.0\} \).

The entries \( X_{ij} \) in the input matrix \( X \) are generated independently through the following process:

1. Sample the true row class \( u_i \in \{1, 2\} \) from the multinomial distribution with probability \( p = (0.3, 0.7) \).

2. Sample the true column class \( v_j \in \{1, 2\} \) from the multinomial distribution with probability \( q = (0.2, 0.8) \).

3. Conditioning on \( u_i \) and \( v_j \), \( X_{ij} \) follows a Gaussian distribution with mean \( M_{u_i v_j} \) and standard deviation \( \Sigma_{u_i v_j} \):

\[
X_{ij}|u_i, v_j \sim \mathcal{N}(M_{u_i v_j}, \Sigma_{u_i v_j}^2). 
\]

Note that \( M \) and \( \Sigma \) are \( 2 \times 2 \) matrices representing the means and standard deviations of entries in different blocks. They are different for each simulation setting.
7.1 Simulation 1: Blocks with Different Means and the Same Variance

In the first simulation, we consider the case where the $2 \times 2$ blocks have different means and the same variance. More specifically, we set

$$M = b \cdot \begin{bmatrix} 0.36 & 0.90 \\ -0.58 & -0.06 \end{bmatrix},$$

where $b \in \{0.20, 0.25, 0.30\}$. The entries of the matrix are simulated from a uniform distribution on $[-1, 1]$. As $b$ increases, the difference between the means in different blocks also increases. In addition, we set

$$\Sigma = \begin{bmatrix} 1 + b & 1 \\ 1 & 1 + b \end{bmatrix},$$

which means that all entries have the same standard deviation of 1. This type of structure is exactly what many biclustering methods including PL and SBC assume the input matrix $X$ has, therefore we would expect their performance to be good.

Table 1: The means (and standard errors) of the misclassification rate for Simulation 1 over 50 simulations.

<table>
<thead>
<tr>
<th>$b$</th>
<th>AKM ($\lambda = 0$)</th>
<th>AKM ($\lambda = 0.1$)</th>
<th>AKM ($\lambda = 1$)</th>
<th>KM</th>
<th>PL</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>$a = 0.5$ 0.722(0.002) 0.591(0.015) 0.527(0.007) 0.542(0.006) 0.328(0.010) 0.409(0.014)</td>
<td>$a = 1.0$ 0.73(0.002) 0.47(0.006) 0.474(0.006) 0.489(0.006) 0.244(0.005) 0.258(0.008)</td>
<td>$a = 2.0$ 0.718(0.002) 0.445(0.006) 0.447(0.006) 0.444(0.006) 0.221(0.004) 0.222(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>$a = 0.5$ 0.715(0.003) 0.457(0.010) 0.450(0.007) 0.467(0.007) 0.201(0.007) 0.223(0.012)</td>
<td>$a = 1.0$ 0.716(0.003) 0.415(0.010) 0.415(0.010) 0.404(0.007) 0.150(0.006) 0.151(0.006)</td>
<td>$a = 2.0$ 0.706(0.002) 0.347(0.013) 0.338(0.013) 0.321(0.006) 0.147(0.004) 0.148(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>$a = 0.5$ 0.711(0.003) 0.392(0.012) 0.387(0.011) 0.386(0.010) 0.110(0.004) 0.110(0.004)</td>
<td>$a = 1.0$ 0.694(0.003) 0.291(0.015) 0.270(0.013) 0.287(0.010) 0.083(0.004) 0.079(0.004)</td>
<td>$a = 2.0$ 0.670(0.004) 0.173(0.007) 0.169(0.005) 0.197(0.007) 0.080(0.003) 0.081(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results are reported in Table 1. Under this setting, we see that PL and SBC have similar misclassification rates, both much smaller than the other four methods. KM, AKM with $\lambda = 1$ and $\lambda = 0.1$ also have similar misclassification rates, though they are significantly larger than PL and SBC. AKM with $\lambda = 0$ has the worst performance, with misclassification rates around 0.7 in all cases. In addition, we observe a general trend that as $a$ and $b$ increase, the misclassification rates decrease. This trend agrees with our expectation, because larger $a$ means larger input matrix, and larger $b$ means larger difference between the means in different blocks, both of which should improve the performance of biclustering methods.

7.2 Simulation 2: Blocks with Different Variances and the Same Mean

In the second simulation, we consider the case where the $2 \times 2$ blocks have different variances and the same mean. More specifically, we set

$$\Sigma = \begin{bmatrix} 1 + b & 1 \\ 1 & 1 + b \end{bmatrix},$$
where \( b \in \{0.20, 0.25, 0.30\} \). As \( b \) increases, the difference between the standard deviations in different blocks also increases. In addition, we set

\[
M = \begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix},
\]

which means that all entries have the same mean of 0. This is the case where the blocks are defined not by different means but by different variances, and many biclustering methods including PL and SBC are unable to detect this type of structure.

<table>
<thead>
<tr>
<th>( a )</th>
<th>( b = 0.20 )</th>
<th>( b = 0.25 )</th>
<th>( b = 0.30 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.475(0.015)</td>
<td>0.372(0.031)</td>
<td>0.099(0.015)</td>
</tr>
<tr>
<td>1.0</td>
<td>0.650(0.013)</td>
<td>0.698(0.003)</td>
<td>0.701(0.003)</td>
</tr>
<tr>
<td>2.0</td>
<td>0.150(0.014)</td>
<td>0.719(0.002)</td>
<td>0.720(0.002)</td>
</tr>
</tbody>
</table>

Table 2: The means (and standard errors) of the misclassification rate for Simulation 2 over 50 simulations.

Results are reported in Table 2. Under this setting, we see that KM, PL, SBC and AKM with \( \lambda = 1 \) all have similarly bad performance, with misclassification rates around 0.72 in all cases. AKM with \( \lambda = 0 \) achieves the smallest misclassification rates in all cases, and in some cases (\( a \in \{1.0, 2.0\}, b \in \{0.25, 0.30\} \)) even produces near perfect biclustering results. AKM with \( \lambda = 0.1 \) also has good performance, with slightly larger misclassification rates compared to AKM with \( \lambda = 0 \). In addition, with regard to AKM with \( \lambda = 0 \) and \( \lambda = 0.1 \), we also see the general trend that as \( a \) and \( b \) increase, the misclassification rates decrease. Interestingly, in this setting larger \( b \) means larger difference between the means in different blocks, and only AKM with \( \lambda = 0 \) and \( \lambda = 0.1 \) are able to detect and respond to this type of structure.

### 7.3 Simulation 3: Blocks with Different Means and Different Variances

In the third simulation, we consider the case where the \( 2 \times 2 \) blocks have different means and different variances, which is a combination of the first and second case. More specifically, we set

\[
M = b \cdot \begin{bmatrix}
0.36 & 0.90 \\
-0.58 & -0.06
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
1 + b & 1 \\
1 & 1 + b
\end{bmatrix},
\]

where \( b \in \{0.20, 0.25, 0.30\} \). As \( b \) increases, the difference between the means and standard deviations in different blocks also increases. This type of structure is arguably the most
common type in practice, where different biclusters not only have different means but also have different variances.

Results are reported in Table 3. Under this setting, we see that KM and AKM with $\lambda = 1$ have similar and the worst performance. PL and SBC also have similar but slightly better performance compared to KM and AKM with $\lambda = 1$. Most importantly, we see again that AKM with $\lambda = 0$ significantly outperforms all other methods in all cases, and has misclassification rates less than 0.05 in all cases except when $a = 0.50$ and $b = 0.20$. In addition, AKM with $\lambda = 0.1$ also performs much better than KM, PL and SBC in all cases except when $a = 0.50$ and $b = 0.20$, with slightly larger misclassification rates compared to AKM with $\lambda = 0$.

Comparing the results in Table 3 to those in Table 1, we see that the additional difference between the variances in different blocks significantly benefits AKM with $\lambda = 0$ and $\lambda = 0.1$, resulting in a drastic decrease in misclassification rates. In contrast, this additional difference between block variances harms the performance of AKM with $\lambda = 1$, KM, PL, and SBC, causing varying degrees of increase in misclassification rates. This indicates that AKM with $\lambda = 0$ and $\lambda = 0.1$ are capable of leveraging the information about difference between block variances to detect meaningful structures, whereas AKM with $\lambda = 1$, KM, PL, and SBC are adversely affected by difference between block variances.

Comparing the results in Table 3 to those in Table 2, we see that the additional difference between the means in different blocks significantly benefits AKM with $\lambda = 1$, KM, PL, and SBC, which is expected because many biclustering methods including PL and SBC make the explicit assumption that different biclusters should have different means. However, this additional difference between block means also benefits AKM with $\lambda = 0$ and $\lambda = 0.1$, resulting in even better performance.

Importantly, in Simulation 3, larger $b$ means larger difference between both the means and the variances in different blocks, so there are two kinds of signals present. In this situation, AKM with $\lambda = 0$ and $\lambda = 0.1$ perform much better than KM, PL and SBC, indicating the possibility that AKM with appropriate $\lambda$ is suitable for dealing with complex data sets in the real world.

<table>
<thead>
<tr>
<th></th>
<th>AKM ($\lambda = 0$)</th>
<th>AKM ($\lambda = 0.1$)</th>
<th>AKM ($\lambda = 1$)</th>
<th>KM</th>
<th>PL</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 0.5$</td>
<td>0.319(0.018)</td>
<td>0.494(0.024)</td>
<td>0.632(0.008)</td>
<td>0.612(0.010)</td>
<td>0.422(0.010)</td>
<td>0.516(0.011)</td>
</tr>
<tr>
<td>$a = 1.0$</td>
<td>0.031(0.004)</td>
<td>0.183(0.015)</td>
<td>0.535(0.006)</td>
<td>0.542(0.004)</td>
<td>0.320(0.005)</td>
<td>0.334(0.008)</td>
</tr>
<tr>
<td>$a = 2.0$</td>
<td>0.009(0.005)</td>
<td>0.143(0.019)</td>
<td>0.483(0.003)</td>
<td>0.479(0.004)</td>
<td>0.279(0.004)</td>
<td>0.278(0.003)</td>
</tr>
</tbody>
</table>

Table 3: The means (and standard errors) of the misclassification rate for Simulation 3 over 50 simulations.
8. Applications

In this section, we apply our algorithm to three cancer gene expression data sets, all of which were proposed and preprocessed by de Souto et al. (2008). In all three data sets, the rows represent different samples of tissues, and the columns represent different genes. The samples have already been classified into different groups based on their types of tissue, which means that the true sample cluster labels are available. This enables us to evaluate and compare the performance of our algorithm with three other biclustering algorithms (KM, PL, SBC) in terms of sample misclassification rate, which is defined as

\[
\text{sample misclassification rate} = \frac{\text{number of samples classified into the wrong cluster}}{\text{total number of samples}}.
\]

Smaller sample misclassification rate indicates better performance at clustering samples.

8.1 Breast and Colon Cancer Gene Expression Data Set

The first data set consists of 104 samples and 182 genes. There are only two types of samples: 62 samples correspond to breast cancer tissues, and 42 samples correspond to colon cancer tissues.

When applying our biclustering algorithm to real-world data, sometimes we do not have prior knowledge about the appropriate number of biclusters \(k\). In that case, one good way to select \(k\) is the “elbow method”, which is also a widely used heuristic method to determine the number of clusters \(k\) in traditional \(k\) means clustering. The idea is to run the algorithm with \(\lambda = 0\) and calculate the loss for different values of \(k\), make a plot with loss on the \(y\)-axis and \(k\) on the \(x\)-axis, and select the \(k\) at the point of inflection (the “elbow” of the curve). In Figure 2, we plot the losses for \(k\) from 1 to 10 when applying our algorithm with \(\lambda = 0\) to the breast and colon cancer gene expression data set. By looking at Figure 2, it is quite clear that our algorithm should select \(k = 2\) as the number of biclusters, which is also the true number of row clusters.

![Figure 2: Losses for different k on the breast and colon cancer gene expression data set.](image)
Having selected $k = 2$ as the number of biclusters, we apply our algorithm with three different $\lambda$ values: 0, 0.1 and 1. For comparison, we also apply $k$-means clustering on the rows (KM), profile likelihood biclustering (PL), and sparse biclustering (SBC), with the number of row clusters set to 2. Both PL and SBC allow the number of column clusters to be different from the number of row clusters, so we vary the number of column clusters from 1 to 20 and report the best result. For PL, the distribution family is selected as Gaussian. For SBC, the input matrix is always mean-centered before applying the method, and the tuning parameter $\lambda$ is selected by choosing the $\lambda$ with the smallest BIC over a grid of $\lambda$ values, both of which are suggested in their paper. All methods are run with 100 random initializations.

<table>
<thead>
<tr>
<th>AKM ($\lambda = 0$)</th>
<th>AKM ($\lambda = 0.1$)</th>
<th>AKM ($\lambda = 1$)</th>
<th>KM</th>
<th>PL</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0385</td>
<td>0.0385</td>
<td>0.0385</td>
<td>0.3462</td>
<td>0.3462</td>
<td>0.3462</td>
</tr>
</tbody>
</table>

Table 4: The sample misclassification rates on the breast and colon cancer gene expression data set.

The sample misclassification rates are reported in Table 4. Noticeably, all three other biclustering methods have the same sample misclassification rate of 0.3462, which is around ten times larger than the sample misclassification rate of 0.0385, achieved by our algorithm with all three different values of $\lambda$. Compared to KM which ignores the interaction between samples and genes, our algorithm successfully leverages information about the interaction to significantly improve sample clustering performance. In contrast, PL and SBC fail to perform better than KM at clustering samples on this data set.

8.2 Brain Cancer Gene Expression Data Set

The second data set consists of 50 samples and 1739 genes. There are three types of samples: 31, 14, and 5 samples correspond to three different types of brain cancer tissues.

Again, we try to use the elbow method to select the appropriate number of biclusters $k$ for our algorithm. In Figure 3, we plot the losses for $k$ from 1 to 10 when applying our algorithm with $\lambda = 0$ to the brain cancer gene expression data set. In this case, it is not completely clear which $k$ we should select, although $k = 2$ or $k = 8$ might be the two most reasonable choices based on the plot alone. However, in some applications we do have prior knowledge about the appropriate number of biclusters $k$. For example, we know $k = 3$ should be the default choice for this data set, because three row clusters could be naturally defined based on the three types of samples. Moreover, even if we select $k = 2$ or $k = 8$ based on the plot, the resulting biclusters could reveal interesting findings about the subgroups of samples or genes in this data set.

Having selected $k = 3$ as the number of biclusters, we again apply our algorithm with three different $\lambda$ values: 0, 0.1 and 1. We also apply KM, PL, and SBC, with the number of row clusters set to 3. All other settings of the biclustering algorithms are the same as in the first application.

The sample misclassification rates are reported in Table 5. In this case, we again see that our algorithm with all three different values of $\lambda$ achieve the smallest sample misclassification.
Biclustering with Alternating K-Means

Figure 3: Losses for different $k$ on the brain cancer gene expression data set.

<table>
<thead>
<tr>
<th>AKM ($\lambda = 0$)</th>
<th>AKM ($\lambda = 0.1$)</th>
<th>AKM ($\lambda = 1$)</th>
<th>KM</th>
<th>PL</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.36</td>
<td>0.34</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 5: The sample misclassification rates on the brain cancer gene expression data set.

rate of 0.22, whereas other three biclustering methods all have sample misclassification rates around 0.34. This result indicates that even on larger gene expression data sets with more than two biclusters, our algorithm is still able to significantly outperform other biclustering methods such as KM, PL and SBC.

8.3 Prostate Cancer Gene Expression Data Set

The third data set consists of 92 samples and 1288 genes. There are four types of samples: 27 samples correspond to benign prostate tissues, and 13, 32, 20 samples correspond to prostate cancer tissues of three different stages, respectively.

Again, we try to use the elbow method to select the appropriate number of biclusters $k$ for our algorithm. In Figure 4, we plot the losses for $k$ from 1 to 10 when applying our algorithm with $\lambda = 0$ to the prostate cancer gene expression data set. In this case, it is clear that our algorithm should select $k = 4$ as the number of biclusters, which is also the true number of row clusters.

Having selected $k = 4$ as the number of biclusters, we again apply our algorithm with three different $\lambda$ values: 0, 0.1 and 1. We also apply KM, PL, and SBC, with the number of row clusters set to 4. All other settings of the biclustering algorithms are the same as in the first and second application.

<table>
<thead>
<tr>
<th>AKM ($\lambda = 0$)</th>
<th>AKM ($\lambda = 0.1$)</th>
<th>AKM ($\lambda = 1$)</th>
<th>KM</th>
<th>PL</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5217</td>
<td>0.4239</td>
<td>0.4239</td>
<td>0.5652</td>
<td>0.5109</td>
<td>0.5543</td>
</tr>
</tbody>
</table>

Table 6: The sample misclassification rates on the prostate cancer gene expression data set.
The sample misclassification rates are reported in Table 6. In this case, we see that AKM with $\lambda = 0.1$ and $\lambda = 1$ achieve the smallest sample misclassification rate of 0.4239. AKM with $\lambda = 0$ has a sample misclassification rate of 0.5217, which is slightly worse than PL but better than SBC and KM. This result once again demonstrates our algorithm’s ability to achieve better performance at clustering samples on gene expression data sets compared to other related biclustering algorithms such as KM, PL, and SBC.

9. Discussion

In this paper, we have provided a new formulation of the biclustering problem based on the idea of minimizing the empirical clustering risk. We have developed and proved a consistency result with respect to the empirical clustering risk. Since the optimization problem is combinatorial in nature, finding the global minimum is computationally infeasible. In light of this fact, we have proposed a simple and novel algorithm that finds a local minimum by alternating the use of $k$-means clustering between columns and rows, and released an R package `akmbiclust` on CRAN that implements the algorithm. We have also provided a probabilistic interpretation of the optimization problem, and proposed extending our method by adding penalization terms. We have evaluated and compared the performance of our algorithm to other related biclustering methods on both simulated data and real-world gene expression data sets. The results have demonstrated that our algorithm is able to detect meaningful structures in the data and outperform other competing biclustering methods in a lot of situations.

One big advantage of our algorithm is its simplicity: the $k$ biclusters can be found simply by applying $k$-means clustering between columns and rows alternately. However, the simplicity comes at the expense of flexibility: by assigning every row and every column to one and only one bicluster, our method excludes the possibility of overlapping biclusters. Although allowing the biclusters to overlap might be a more reasonable assumption in some cases, we argue that trading off some flexibility for more simplicity is a worthwhile choice for many applications.
In the future, we plan to explore a more general setting of biclustering: biclustering on graphs. The idea is that each column of $X$ represents a vertex in a graph $G$, and each row of $X$ represents a measurement on the vertices. The graph structure of $G$ imposes some restrictions on the column partitions, namely the columns in every bicluster should correspond to vertices in a connected subgraph of $G$. The problem formulation in Section 2 can be considered as the special case where there is no restriction on the column partitions, and that is equivalent to the graph $G$ being a complete graph with $m$ vertices.

Under this setting of biclustering on graphs, a consistency result could be developed and proved in a very similar way. The only difference would be that the optimization would be over all possible choices of column partitions that “preserve” the graph structure of $G$. All three lemmas and the main theorem still hold true after adding the requirement of $I \in \mathcal{I}(G)$, where $\mathcal{I}(G)$ denote the set of column partitions that “preserve” the graph structure of $G$. However, the algorithm presented in Section 5 no longer applies to this setting, and in general people need to either do exhaustive searches over all column partitions in the space of $\mathcal{I}(G)$, or find some heuristic method that could efficiently search through the space of $\mathcal{I}(G)$. This would depend on the specific graph structure of $G$.

References


