Biclustering of Binary Data Matrices: The BiBit Workflow

Proof of Concept

July 20, 2018

1 Application of the BiBit Workflow on Generated Scenarios

1.1 Data Generation

In order to demonstrate the effectiveness of the BiBit Workflow, we will generate the following four data sets. These data sets contain multiple biclusters with or without noise. Further, due to the column similarity feature of the workflow, overlap between the bicluster columns was also considered in these data sets.

	Data 1	**Data 2**	**Data 3**	**Data 4**
Data Matrix	10000 × 50	10000 × 50	10000 × 50	10000 × 50
Background Signal	15%	15%	15%	15%
Biclusters	BC1: 200 × 5	BC1: 200 × 5	BC1: 200 × 10	BC1: 200 × 10
	BC2: 100 × 5	BC2: 100 × 5	BC2: 100 × 10	BC2: 100 × 10
	BC3: 200 × 10	BC3: 200 × 10	BC3: 200 × 10	BC3: 200 × 10
	BC4: 100 × 10	BC4: 100 × 10	BC4: 100 × 10	BC4: 100 × 10
Overlap	None	None	BC1∩BC2: 5 columns	BC1∩BC2: 5 columns
			BC3∩BC4: 2 columns	BC3∩BC4: 2 columns
BC Signal	100%	90%	100%	90%

Table 1: Generated data sets.

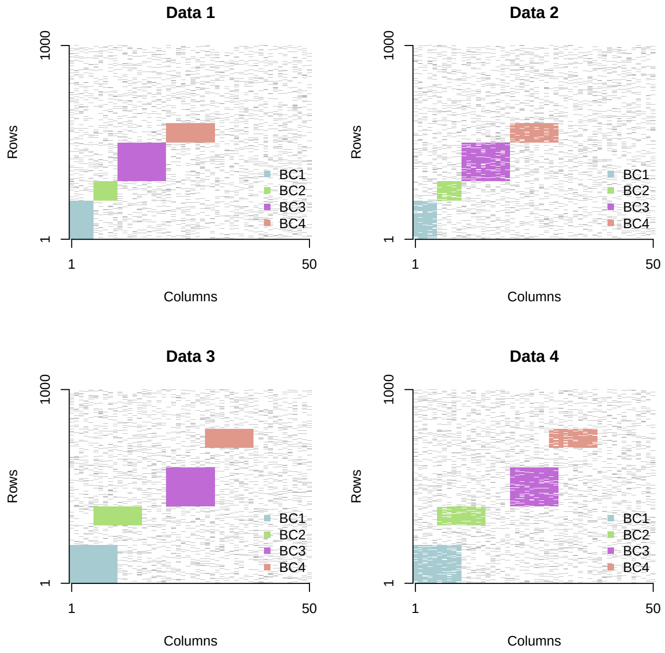


Figure 1: Visualisation of the four generated data sets, unshuffled and subsetted on the first 1000 rows.

	Data 1				**Data 2**				**Data 3**				**Data 4**			
Row Noise	BC1	BC2	BC3	BC4	BC1	BC2	BC3	BC4	BC1	BC2	BC3	BC4	BC1	BC2	BC3	BC4
0	200	100	200	100	121	67	66	32	200	100	200	100	72	24	70	28
1	0	0	0	0	59	23	83	43	0	0	0	0	78	50	79	43
2	0	0	0	0	15	10	34	17	0	0	0	0	33	18	39	25
3	0	0	0	0	5	0	10	6	0	0	0	0	16	6	11	4
4	0	0	0	0	0	0	7	2	0	0	0	0	1	2	1	0

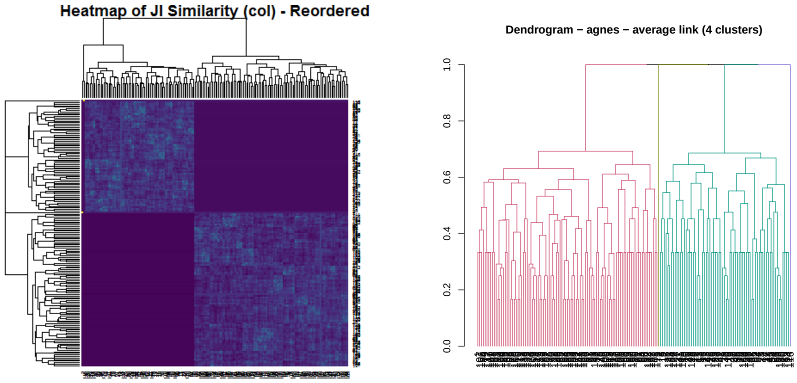
Table 2: Number of rows of 0, 1, 2, 3 and 4 noise levels in generated biclusters.

1.2 Results

The results of the four generated data sets will now be presented following the BiBit Workflow steps described in Section **??**.

Data 1: No Overlap & No Noise

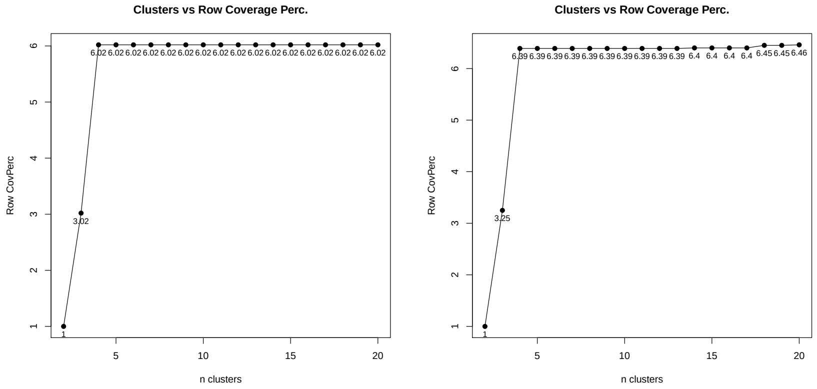
In the first step the original bibit is applied with minimum bicluster dimensions of 50 × 5 which results in 172 biclusters. Next, the bicluster column similarity is determined and clustered in Figure 2.



(a) (b)

Figure 2: Results of Data 1 for step 2 and 3. Panel a: reordered heatmap of column similarity between the biclusters. Panel b: dendrogram of column similarity using hierarchical clustering with average link.

Due to the two singletons we find in Figure 2b, the gap statistic (Tibs2001SEmax) proposes selecting two clusters. However, investigating the row coverage plot in Figure 3 shows that selecting four clusters might be more appropriate.



(a) (b)

Figure 3: Row coverage plot for Data 1 up to 20 clusters. Panel a: zero noise level. Panel b: 10% noise level.

Growing rows on these four merged column patterns results in the biclusters given in Table 3. The biclusters will have a different number of rows depending on the requested noise level. When looking for the simulated perfect biclusters, we find that they they are fully discovered in Table 3a with two additional rows for BC 2.

	BC1	BC2	BC3	BC4
Number of Rows	100	202	200	100
Number of Columns	10	5	10	5

(a)

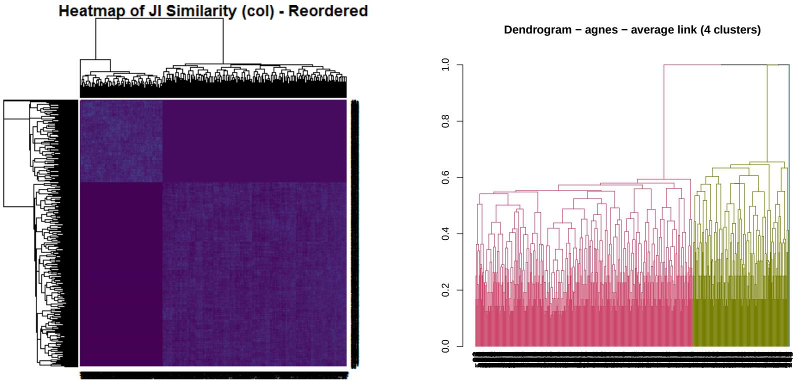
	BC1	BC2	BC3	BC4
Number of Rows	100	219	200	123
Number of Columns	10	5	10	5

(b)

Table 3: Table of final bibit workflow result of data 2. Panel a: for a zero noise level. Panel b: for a noise level of 10% which allows a single zero in each row (due to rounding up).

Data 2: No Overlap & Noise

Applying BiBit with minimum bicluster dimensions of 50 × 5 results in 671 biclusters. Next, the clustered column similarity is shown in Figure 4.

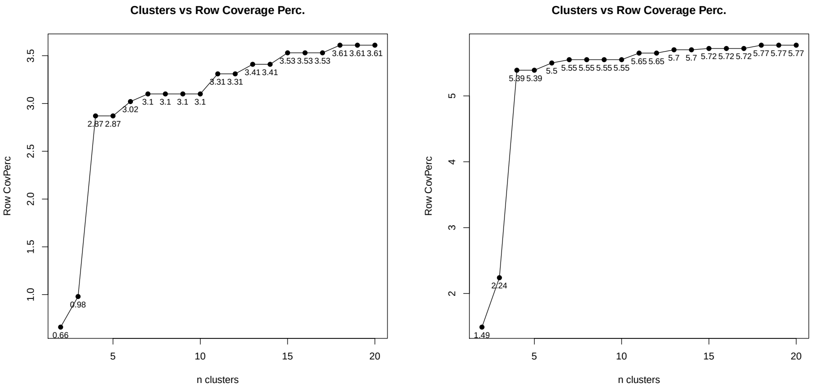


(a)

(b)

Figure 4: Results of Data 2 for step 2 and 3. Panel a: reordered heatmap of column similarity between the biclusters. Panel b: dendrogram of column similarity using hierarchical clustering with average link.

Again due to the two singletons (see Figure 4b), the gap statistic (Tibs2001SEmax) suggests two clusters. However similar to the result of data 1, the row coverage plot seems to indicate 4 clusters (Figure 5).



(a)

(b)

Figure 5: Row coverage plot for Data 2 up to 20 clusters. Panel a: zero noise level. Panel b: single noise level.

Growing rows on these four merged column patterns results in the biclusters given in Table 4. We discover the correct column patterns from the simulated biclusters. Retrieving the correct rows depends on how high the noise level is put in the analysis. The higher the noise level, the more rows are added to the patterns, and the more noisy rows we find of the true bicluster (see Table 2). However note that for BC 1, 3 and 4 that there are a large amount of other rows which also fit this pattern with the requested noise level.

	BC1	BC2	BC3	BC4
Number of Rows	66	32	122	67
Number of Columns	10	10	5	5

(a)

	BC1	BC2	BC3	BC4
Number of Rows	149	75	205	112
Number of Columns	10	10	5	5

(b)

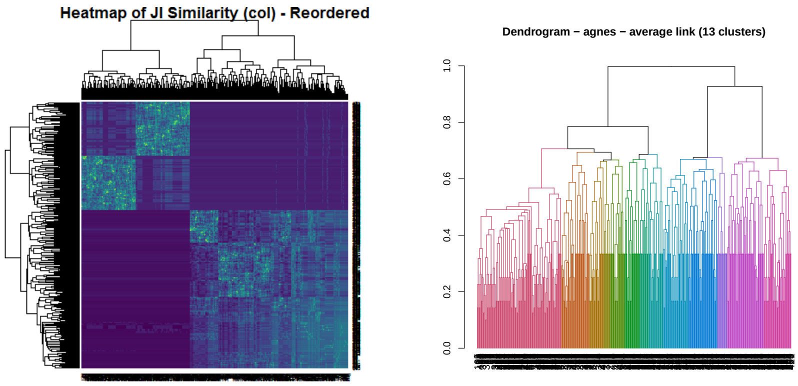
	BC1	BC2	BC3	BC4
Number of Rows	183	92	437	369
Number of Columns	10	10	5	5

(c)

Table 4: Table of final bibit workflow result of data 2. Panel a: for a zero noise level. Panel b: for a noise level of 1. Panel c: For a noise level of 2.

Data 3: Overlap & No Noise

Applying BiBit with minimum bicluster dimensions of 50 × 5 results in 481 biclusters. Next, the clustered column similarity is shown in Figure 6.



Heatmap of JI Similarity (col) - Reordered

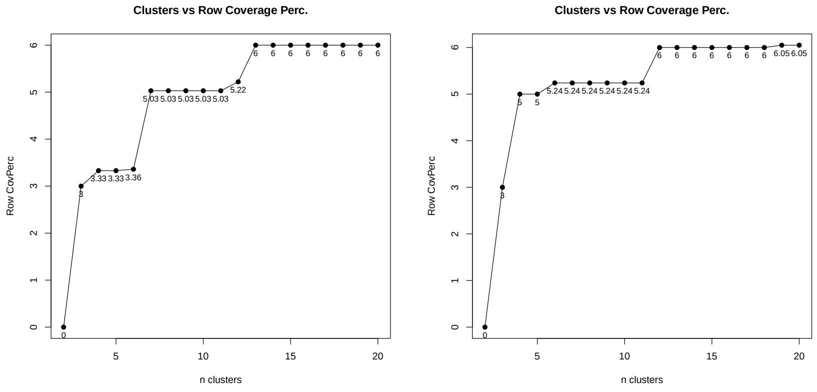
Dendrogram − agnes − average link (13 clusters)

(a)

(b)

Figure 6: Results of Data 3 for step 2 and 3. Panel a: reordered heatmap of column similarity between the biclusters. Panel b: dendrogram of column similarity using hierarchical clustering with average link.

For this data set the gap statistic proposes to select 17 clusters. Depending on the selected noise level this reduces to 8 to 10 biclusters which contain the 4 generated biclusters. Observing the row coverage plots in Figure 7 we could once again go for 4 clusters. However, depending on the allowed noise, we will miss one or two biclusters, namely BC 1 and BC 2 which had 50% column overlap. More specifically, choosing 4 clusters with a zero noise level makes us miss BC 1 and BC 2 for the most part. However when allowing a single noise level, it is only BC 2 that is not discovered. The reason behind this is that the discovered bicluster has 11 columns (including the 10 true ones), so in order to match the true generated bicluster rows, we would need to allow some error. However instead, a higher number of clusters is selected based on the row coverage plot in Figure 7. In Figure 7a the row coverage is shown without allowing any noise in which 13 clusters seem appropriate. The other plot, Figure 7b, allows a single zero and seems to suggest 12 clusters.



(a) (b)

Figure 7: Row coverage plot for Data 3 up to 20 clusters. Panel a: zero noise level. Panel b: single noise level.

Since we are looking for perfect biclusters, selecting 12 clusters for a noise of 0 seems like a good initial choice. Table 5a shows the resulting 8 biclusters. The generated BC 3 and BC 4 which only had a little overlap are easily identified and correspond with BC8 and BC7 in this table, respectively. The generated BC 1 is also discovered as BC5 in this table. BC6 (with its 11 columns) also includes all the correct columns of this same generated bicluster but not its rows. However the generated BC 2 which had 50% overlap with the generated BC 1 is not completely discovered in this result in Table 5a. BC2 misidentifies a single column and cannot match the correct rows. BC1 on the other hand discovers the true columns, but due to the extra columns is also unable to match the true rows.

In order to try and deal with this overlapping issue, we can start to allow some error in the biclusters in order to try and capture all the true biclusters. This result can be found in Table 5b where 12 clusters were chosen with a noise level of 1 which resulted in 6 biclusters. In this result we do identify the true generated biclusters in BC4, BC2, BC6 and BC5, respectively. Note that the only reason we discover the generated BC 2 in the discovered BC2, is because we allow noise. Namely this bicluster contains 11 columns and without error allowance the true rows would not have matched this column pattern.

Finally it should be noted that if 13 clusters would have been chosen for a noise level of 1, the result would not have been this optimal (see Table 5c). Here the generated biclusters 1, 3 and 4 are discovered correctly in BC5, BC7 and BC6, respectively. However the second generated bicluster once again poses some difficulty. Even though all rows of this bicluster are found in BC2 and BC4, both of them miss a single column. And while BC1 contains all columns (and more), here not all rows are identified. What happened here is that going from Table 5c to Table 5b (from 13 to 12 clusters), two patterns seem to have merged together which allowed a better identification of the second generated bicluster.

	BC1	BC2	BC3	BC4	BC5	BC6	BC7	BC8
Number of Rows	3	17	2	100	200	33	100	200
Number of Columns	12	10	12	9	10	11	10	10

(a)

	BC1	BC2	BC3	BC4	BC5	BC6
Number of Rows	24	101	22	200	100	200
Number of Columns	12	11	12	11	10	10

(b)

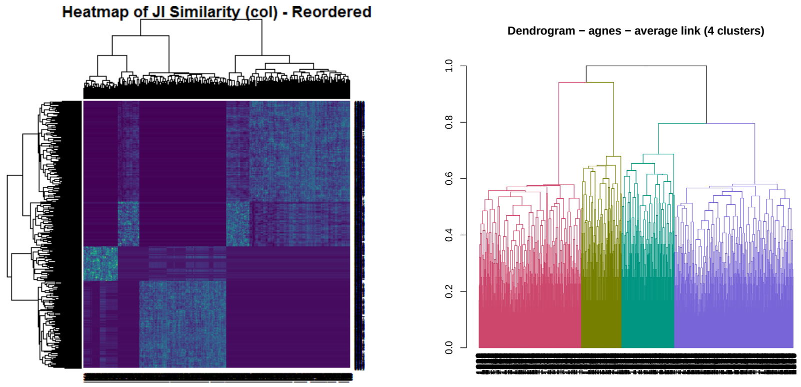
	BC1	BC2	BC3	BC4	BC5	BC6	BC7
Number of Rows	24	101	22	102	200	100	200
Number of Columns	12	10	12	9	11	10	10

(c)

Table 5: Table of final bibit workflow result of data 3. Panel a: 13 clusters for a zero noise level. Panel b: 12 clusters for a noise level of 1. Panel c: 13 clusters for a noise level of 1.

Data 4: Overlap & Noise

For the fourth data set we also apply the original BiBit with minimum bicluster dimensions of 50 × 25. This results in a large number of biclusters, namely 1381. The column similarity is shown in Figure 8

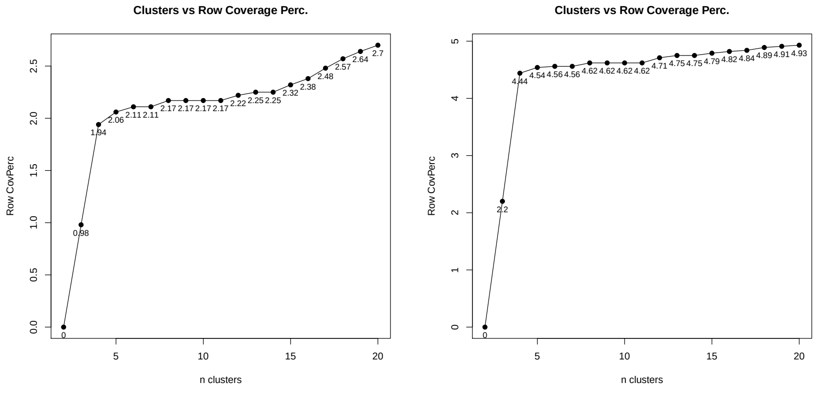


(a)

(b)

Figure 8: Results of Data 4 for step 2 and 3. Panel a: reordered heatmap of column similarity between the biclusters. Panel b: dendrogram of column similarity using hierarchical clustering with average link.

In this example, both the gap statistic (Tibs2001SEmax) and the row coverage plots (Figure 9 suggest 4 clusters.



(a)

(b)

Figure 9: Row coverage plot for Data 4 up to 20 clusters. Panel a: zero noise level. Panel b: single noise level.

Next, using these four merged column patterns, rows are grown condition on a noise level. These results are shown in Table 6 with a noise allowance of 0, 1 and 2. In these results, we perfectly discover the column patterns (even with the overlap), and finding the correct rows is simply a matter of increasing the noise level high enough.

	BC1	BC2	BC3	BC4
Number of Rows	70	28	24	72
Number of Columns	10	10	10	10

(a)

	BC1	BC2	BC3	BC4
Number of Rows	149	71	74	150
Number of Columns	10	10	10	10

(b)

	BC1	BC2	BC3	BC4
Number of Rows	188	96	98	184
Number of Columns	10	10	10	10

(c)

Table 6: Table of final bibit workflow result of data 4. Panel a: for a zero noise level. Panel b: for a noise level of 1. Panel c: for a noise level of 2.

1.3 Discussion

It is generally feasible to retrieve the simulated bicluster examples, even with column overlap. Discovering the true biclusters seems to be a matter of playing with the number of cluster choice and the allowed noise level. This is also something that was found during exploring the case studies where other interesting structures arose when playing with different parameter settings. Further, these examples imply that it is a bit easier to correctly discover the noisy biclusters in contrast to the perfect ones. For overlapping (column) perfect biclusters, it seems to be advisable to allow a certain degree of noise to be able to capture these structures in the data.

Note that if we would have used automatic noise selection, similar results would have been obtained. Since the automatic noise selection always at least allows a single zero, the bicluster discover in data 3 would have been a bit more automated or straight-forward. Finally, reducing the initial bicluster dimensions too much also has an adverse effect on the workflow analysis. For example cutting the minimum row size in two to 25 rows results in a lot of initial small biclusters (and column patterns) which do not merge nicely. Therefore, for larger datas ets it is advised to not start with too small initial bicluster dimensions, both for this reason and for the reason that otherwise BiBit will identify many more biclusters.