Some Bayesian Biclustering Methods: Modeling and Inference

Abhishek Chakraborty and Stephen B. Vardeman

This document supports the findings and analyses in the original research article.

A: Simulation Studies (Section 6)

The tables and plots supporting the simulation studies for the complete dataset scenario are presented below.

1. Analysis of the performance of the algorithm on simulated datasets (6.1)

(i) Biclustering performance

Table 1: Rand index summaries computed between the "true" row clusterings and the three representative row clusterings for each data generating scenario.

	50×60			100×60			100×180		
	SSE	avgRI	lp	SSE	avgRI	lp	SSE	avgRI	lp
Min.	0.85	1	0.49	0.89	0.90	0.75	0.88	0.93	0.56
1^{st} Q.	0.89	1	0.51	0.89	0.90	0.75	0.92	0.95	0.82
Median	0.93	1	0.80	0.92	0.93	0.79	0.94	0.97	0.86
Mean	0.93	1	0.70	0.92	0.93	0.80	0.94	0.96	0.85
3^{rd} Q.	0.99	1	0.85	0.93	0.95	0.84	0.96	0.97	0.94
Max.	1	1	0.88	0.97	0.98	0.89	0.99	0.98	0.98
Std. dev.	0.06	0	0.18	0.03	0.03	0.05	0.03	0.02	0.12

Table 2: Rand index summaries computed between the "true" column clusterings and the three representative column clusterings for each data generating scenario.

	50×60			100×60			100×180		
	SSE	avgRI	lp	SSE	avgRI	lp	SSE	avgRI	lp
Min.	0.76	0.77	0.59	0.90	0.77	0.60	0.82	0.83	0.46
1^{st} Q.	0.89	1	0.61	0.91	1	0.62	0.87	0.83	0.61
Median	0.92	1	0.76	0.96	1	0.74	0.89	0.84	0.62
Mean	0.91	0.95	0.74	0.95	0.98	0.77	0.88	0.87	0.66
3^{rd} Q.	0.95	1	0.87	1	1	0.94	0.90	0.91	0.76
Max.	1	1	0.91	1	1	0.96	0.94	0.93	0.92
Std. dev.	0.08	0.1	0.13	0.05	0.07	0.16	0.03	0.04	0.14

Table 3: Rand index summaries computed between the "true" element-wise clusterings and the three representative element-wise clusterings for each data generating scenario.

	50×60			100×60			100×180		
	SSE	avgRI	lp	SSE	avgRI	lp	SSE	avgRI	lp
Min.	0.86	0.89	0.80	0.95	0.91	0.91	0.96	0.76	0.52
1^{st} Q.	0.92	1	0.81	0.95	0.97	0.91	0.97	0.90	0.77
Median	0.94	1	0.87	0.96	0.98	0.92	0.97	0.97	0.79
Mean	0.93	0.98	0.85	0.96	0.97	0.93	0.97	0.94	0.80
3^{rd} Q.	0.96	1	0.89	0.97	0.99	0.95	0.97	0.99	0.90
Max.	0.98	1	0.92	0.98	0.99	0.95	0.98	1	0.94
Std. dev.	0.04	0.05	0.05	0.01	0.02	0.02	0.01	0.07	0.12

(ii) Performance on a simulated dataset

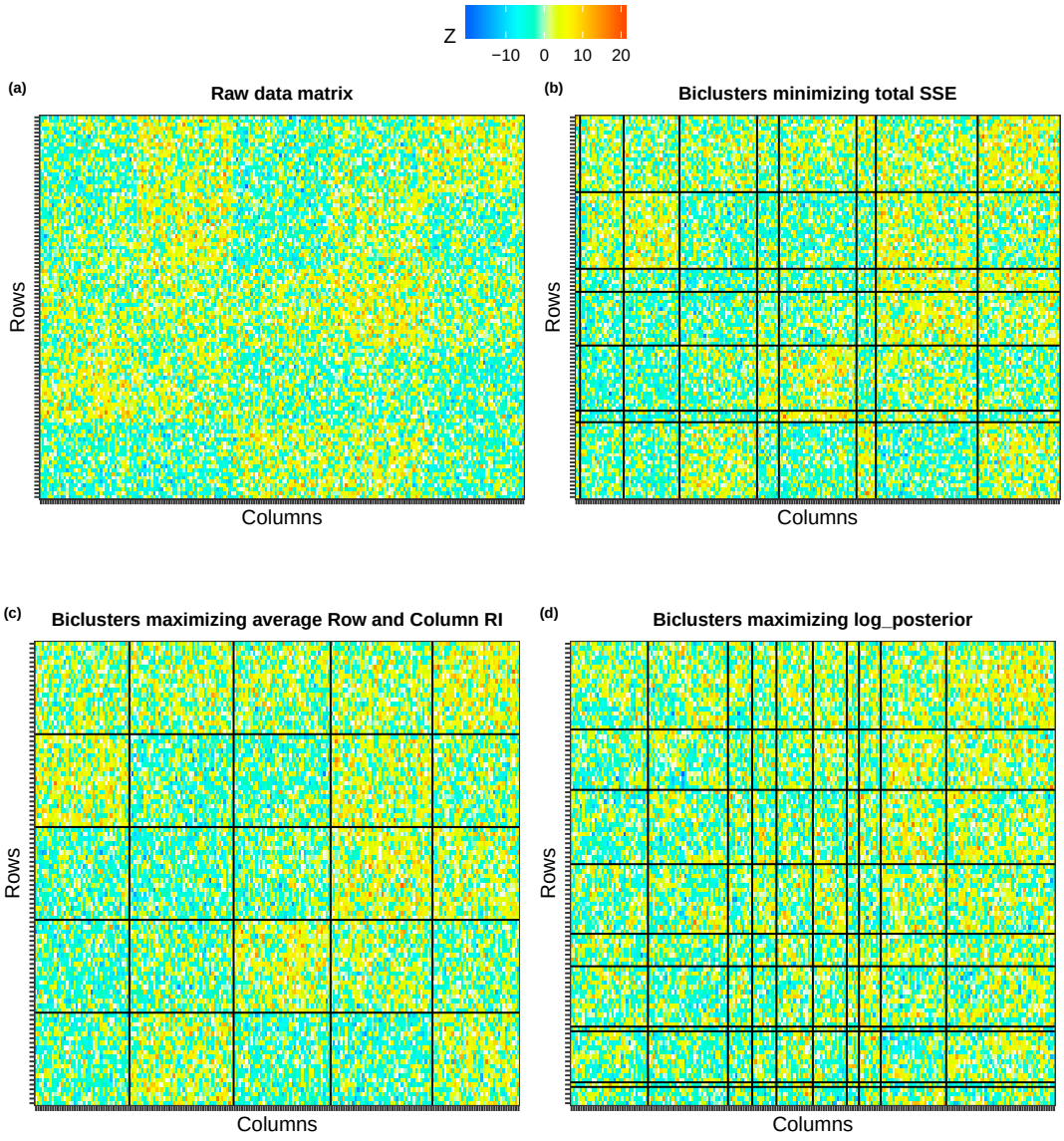

Figure 1: (a) The 100×200 data matrix $\boldsymbol{Y}$ containing 5 row clusters and 5 column clusters. The bicluster patterns identified by the algorithm according to (b) the minimum total SSE criterion, (c) the highest average Rand index criterion, and (d) the maximum log posterior criterion.

Table 4: A summary of total SSE, log posterior, and penalized SSE values computed across $T = 10,000$ complete MCMC iterations and reported against each distinct combination of numbers of non-empty row and column clusters denoted by p and q respectively. Marked in bold are values of the minimum total SSE, the maximum log posterior, the minimum penalized SSE, their corresponding (p, q) combination, and the (p, q) combination having the respective highest average Rand index across all iterates.

p	q	freq	mean SSE	min SSE	max SSE	mean LogPost	min LogPost	max LogPost	min pSSE
5	**5**	**9037**	495347.9	494786.5	496617.6	492.58	459.84	515.06	**30.694**
5	6	800	494849.1	493887.9	495951.7	495.38	466.35	521.10	35.694
5	7	26	494084.3	493409.4	494956	494.82	472.64	515.81	40.693
6	5	97	495001.4	494206.8	495783.8	500.81	479.67	517.54	35.694
6	6	18	494227.1	493665.4	494914.8	501.04	475.26	517.43	41.693
7	7	1	493115.5	493115.5	493115.5	497.08	497.08	497.08	54.693
7	8	7	493053.7	492672	493578.9	529.82	514.67	541.16	61.693
7	**9**	9	492561.7	**491778.4**	493369.6	521.71	506.06	537.27	68.692
8	10	1	503018.2	503018.2	503018.2	530.38	530.38	530.38	85.702
9	10	2	518819.9	513437.1	524202.7	562.77	560.87	564.66	95.710
10	**10**	2	535859.9	530451.7	541268.1	568.21	558.90	**577.51**	105.725

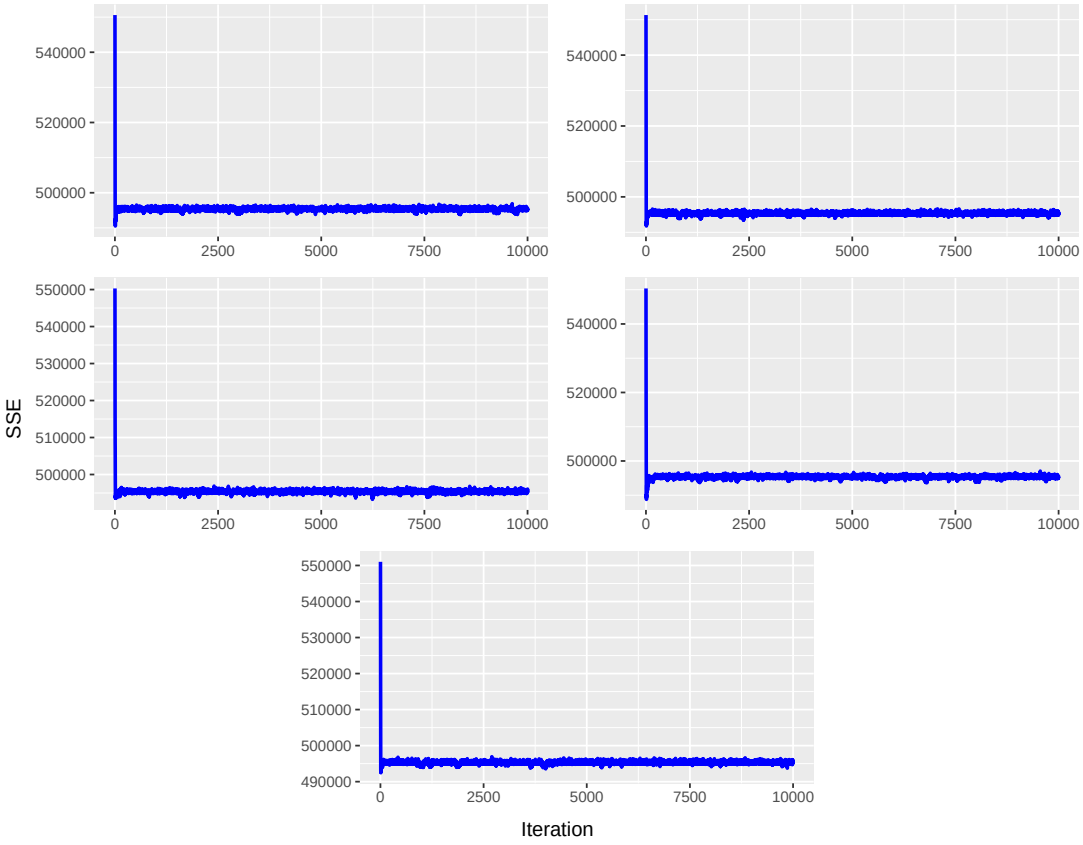

Figure 2: The total SSE plots for the five runs of the algorithm from different starting points.

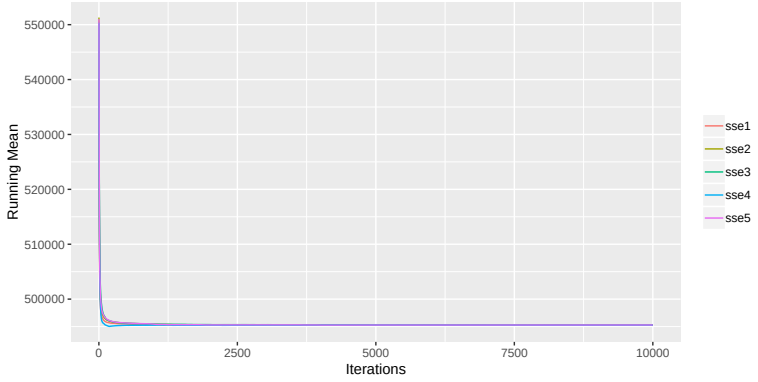

Figure 3: Running mean plot of total SSE values for the five runs of the algorithm with different starting points. The five runs are color-coded.

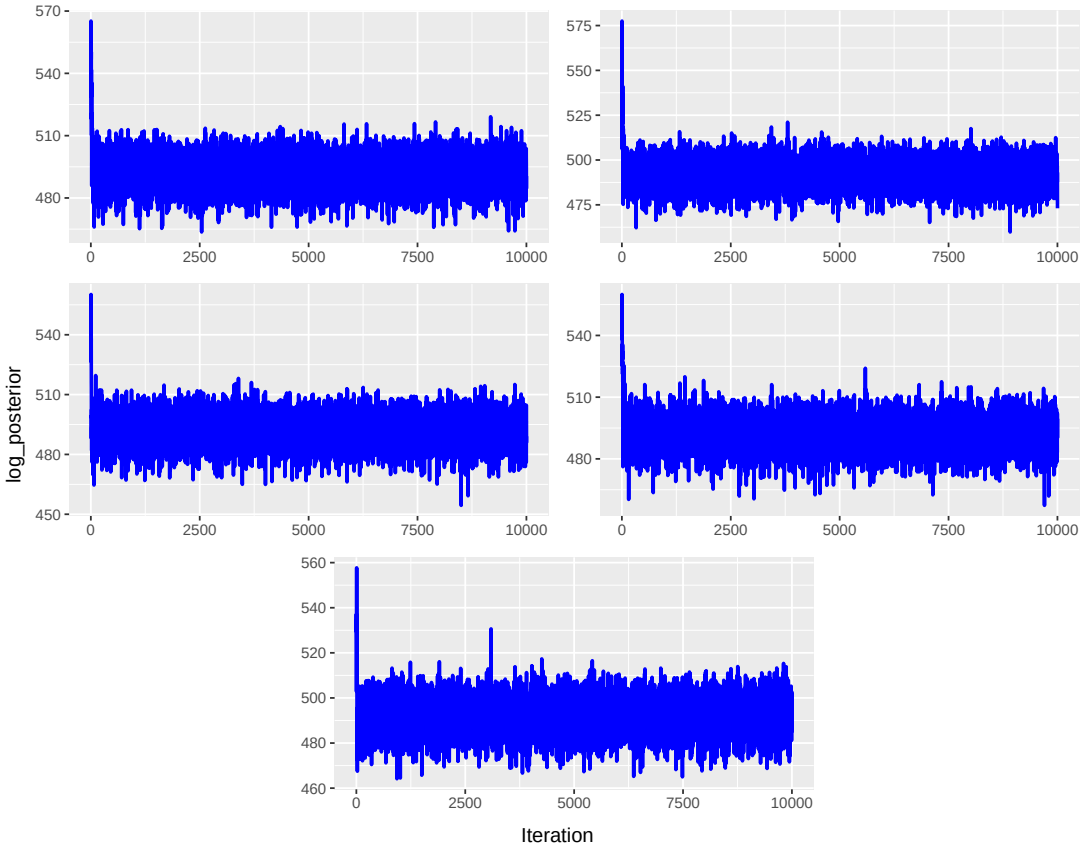

Figure 4: The plots of log posterior values for five runs of the algorithm starting at different points.

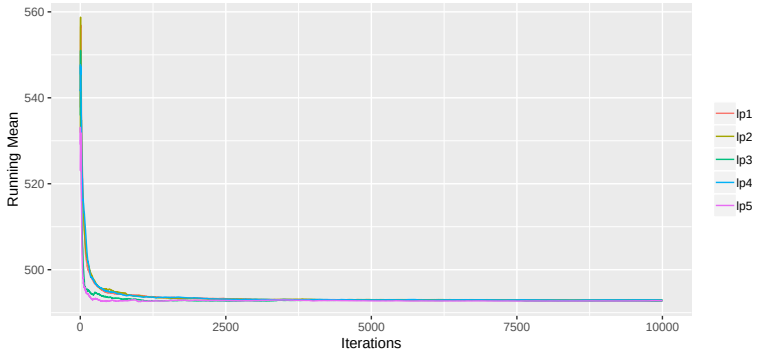

Figure 5: Running mean plot of log posterior values for the five runs of the algorithm with different starting points. The five runs are color-coded.

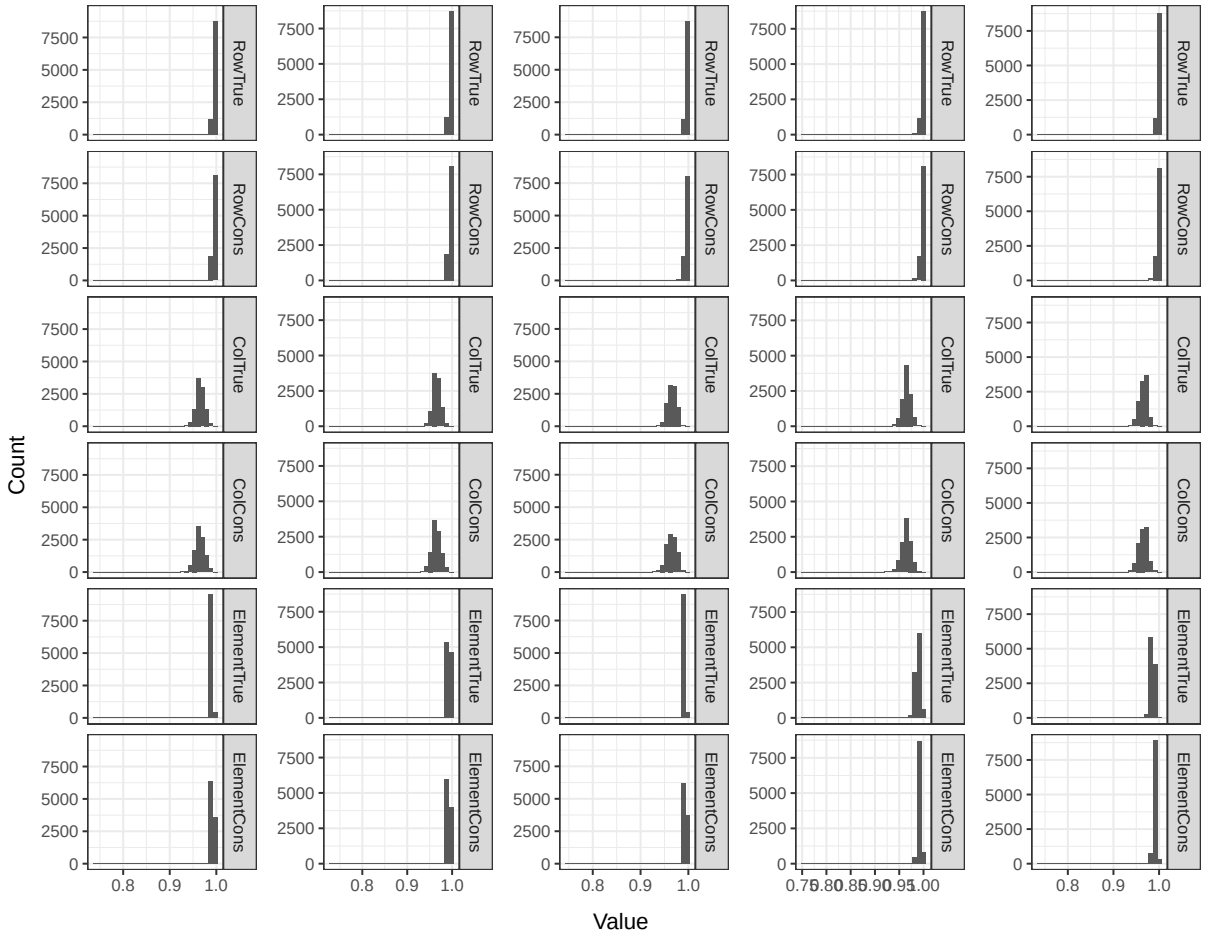

Figure 6: The histograms of Rand index values computed for each run of the algorithm starting at different points. The histograms labelled 'RowTrue', 'ColTrue', and 'ElementTrue' correspond to Rand indices computed between the row, column, and element-wise clustering obtained at each iteration and the "true" row, column, and element-wise partition respectively. The histograms labelled 'RowCons', 'ColCons', and 'ElementCons' correspond to Rand indices computed between two successive row, column, and element-wise partitions respectively.

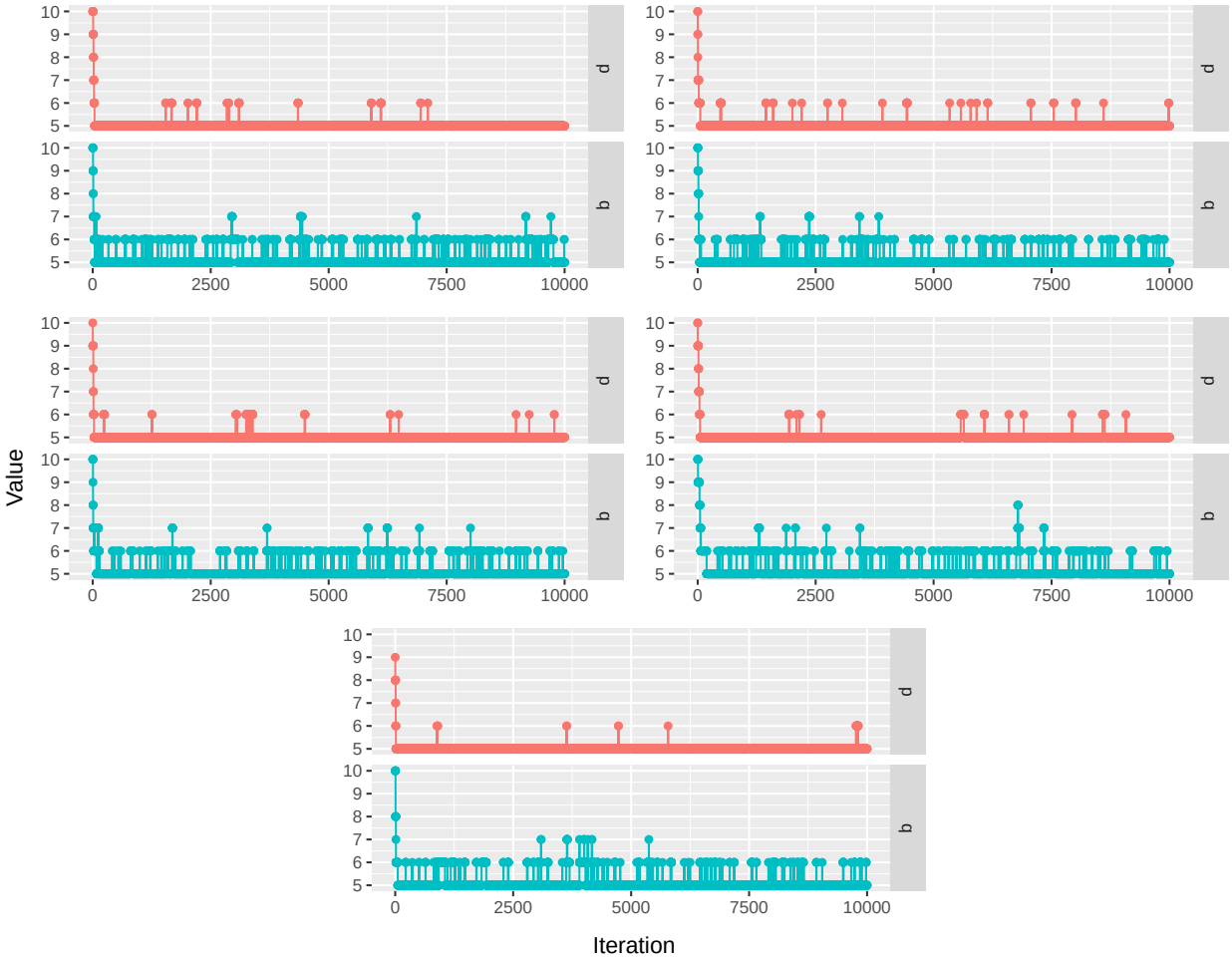

Figure 7: For each run of the algorithm, the numbers of non-empty row and column clusters, denoted by p and q respectively, are plotted against the corresponding iterate.

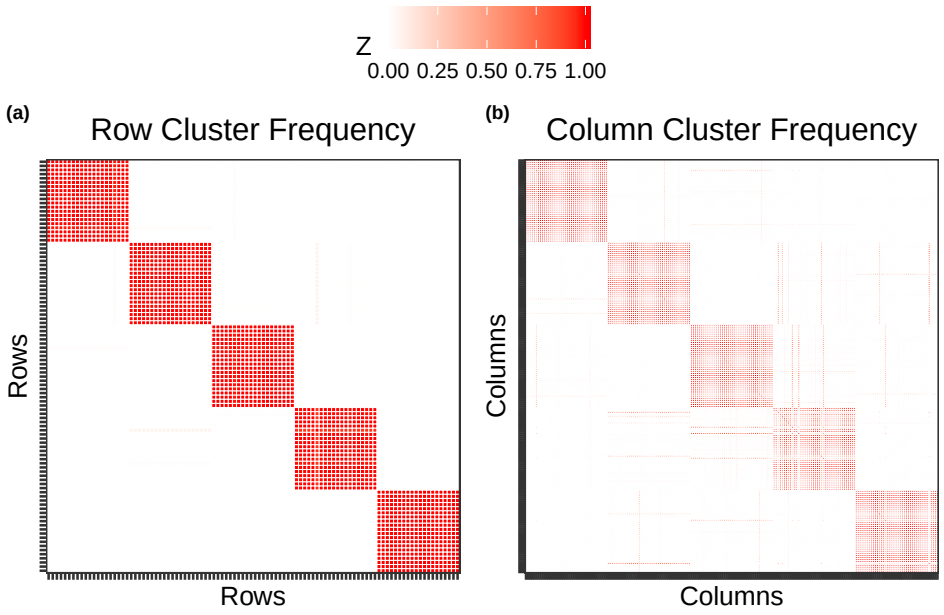

Figure 8: For a particular run of the algorithm: (a) the cluster frequency for the 100 rows, (b) the cluster frequency for the 200 columns in the data matrix $\boldsymbol{Y}$.

2. Comparison with independent one-way k-means clustering (6.2)

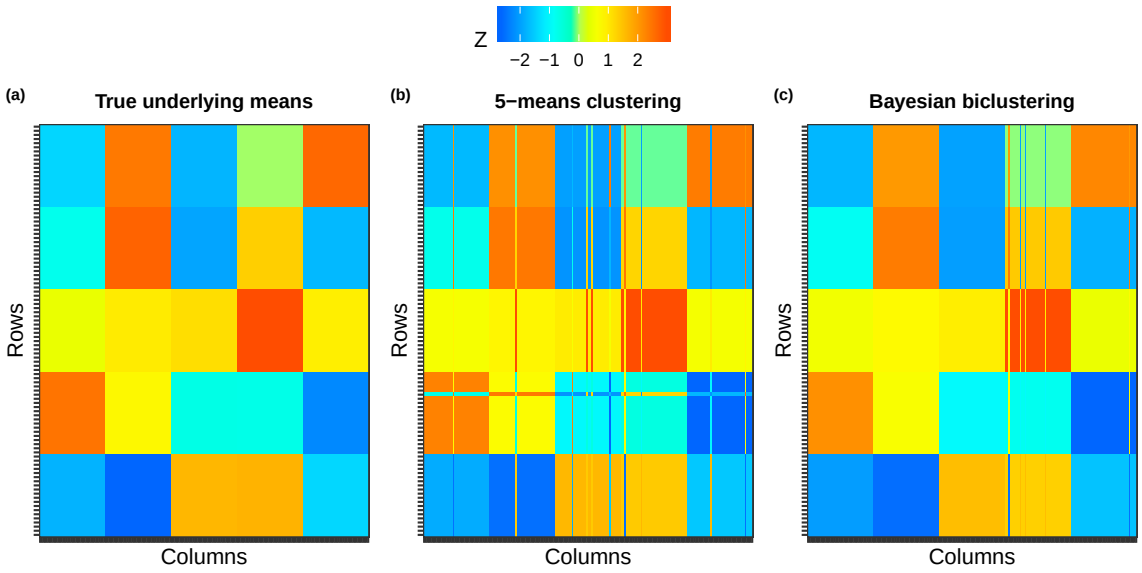

Figure 9: (a) The true underlying means matrix used to generate the 100×200 data matrix $\boldsymbol{Y}$ with 5 row clusters and 5 column clusters. (b) The means matrix estimated by independent one-way 5-means clustering of the rows and columns. (c) The means matrix estimated by the proposed biclustering algorithm. In figures (b) and (c), the incorrectly clustered rows and columns are represented by the horizontal and vertical lines respectively with contradicting colors.

3. Comparison with sparse biclustering (6.3)

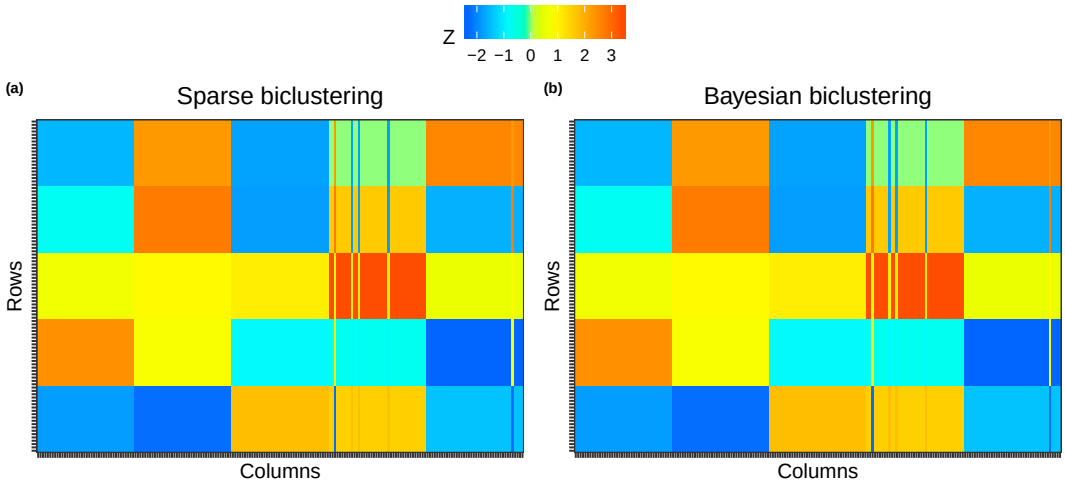

Figure 10: (a) The means matrix estimated by the sparse biclustering technique. (b) The means matrix estimated by our proposed methodology. In figures (a) and (b), the incorrectly clustered rows and columns are represented by the horizontal and vertical lines respectively with contradicting colors.

4. Multiplicative biclusters (6.4)

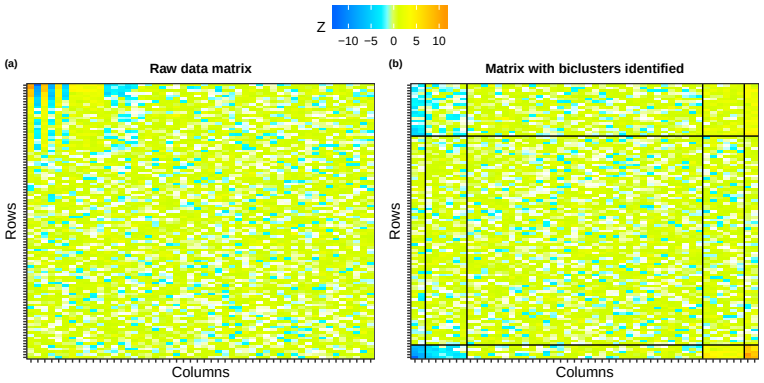

Figure 11: (a) Heatmap of the simulated 100×50 data matrix $\boldsymbol{Y}$ generated from an underlying means matrix $\boldsymbol{M}$ containing multiplicative biclusters. (b) The biclusters identified by the proposed biclustering algorithm according to the highest average Rand index criterion.

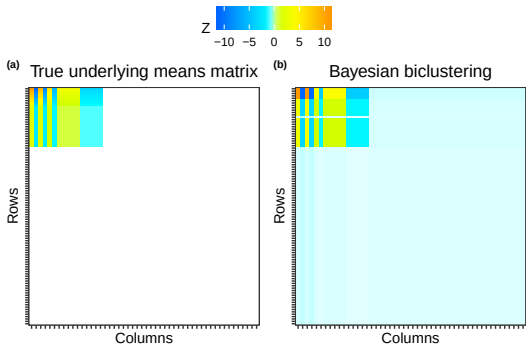

Figure 12: (a) The true underlying means matrix $\boldsymbol{M}$ used to generate the data matrix $\boldsymbol{Y}$. (b) The means matrix estimated by the proposed biclustering algorithm.

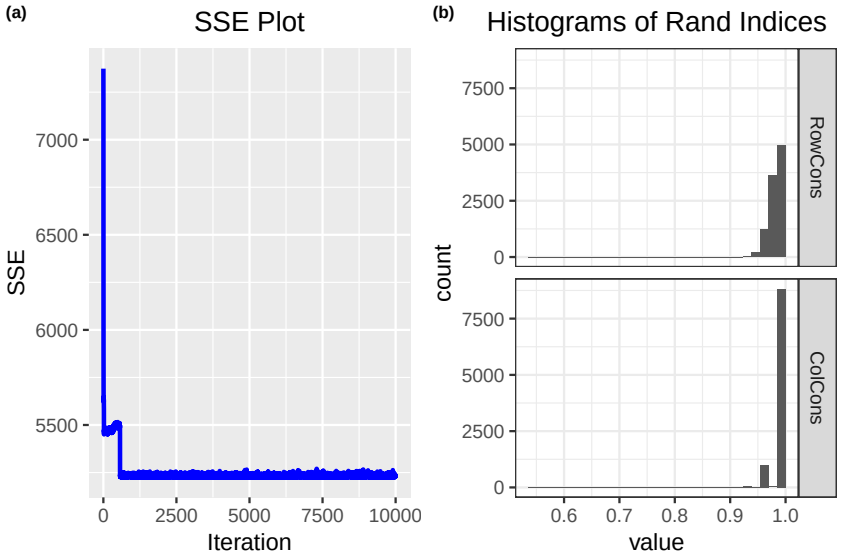

Figure 13: For simulation study 6.4: (a) the total SSE plotted against iterations, (b) the histograms of row and column Rand indices computed between two successive iterates.

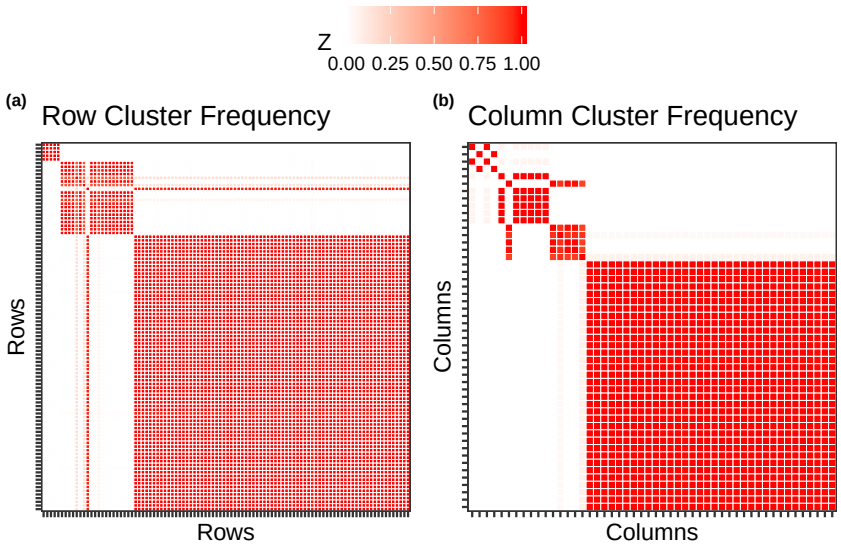

Figure 14: The row and column cluster frequency plots from simulation study 6.4.

5. Overlapping multiplicative biclusters (6.5)

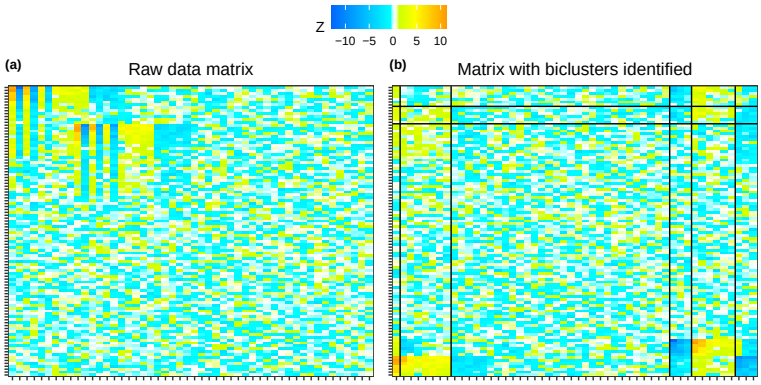

Figure 15: (a) Heatmap of the 100×50 data matrix $\boldsymbol{Y}$ generated from an underlying means matrix $\boldsymbol{M}$ containing overlapping multiplicative biclusters. (b) The biclusters identified by the proposed biclustering algorithm according to the highest average Rand index criterion.

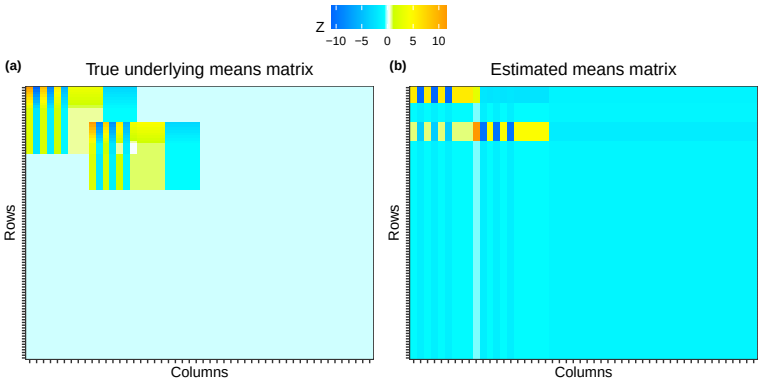

Figure 16: (a) The true underlying means matrix $\boldsymbol{M}$ used to generate the data matrix $\boldsymbol{Y}$. (b) The means matrix estimated by the proposed biclustering algorithm.

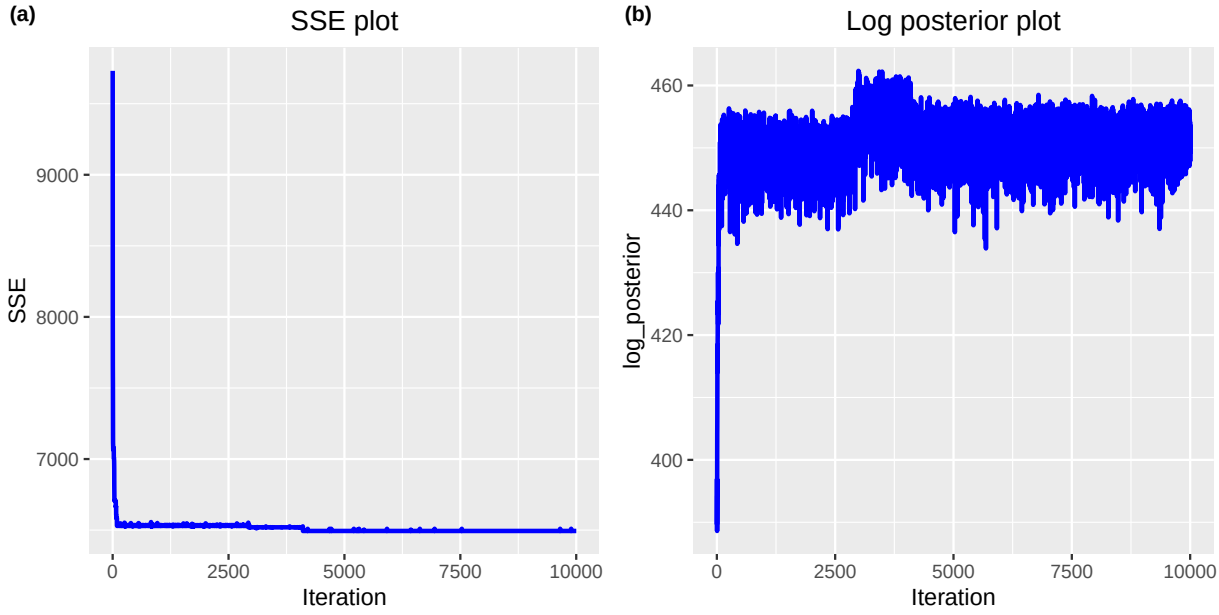

Figure 17: For simulation study 6.5: (a) the total SSE, (b) the log posterior plotted against iterations.

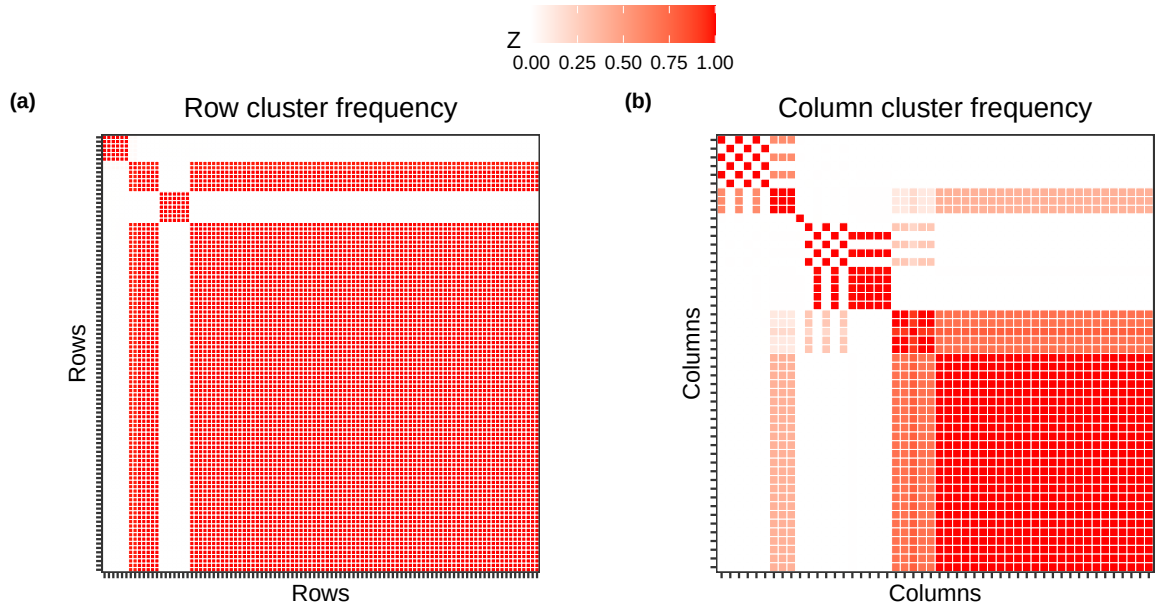

Figure 18: The row and column cluster frequency plots.

B: Application to a real dataset (Section 7)

The following plots support the analysis of the lung cancer gene expression dataset.

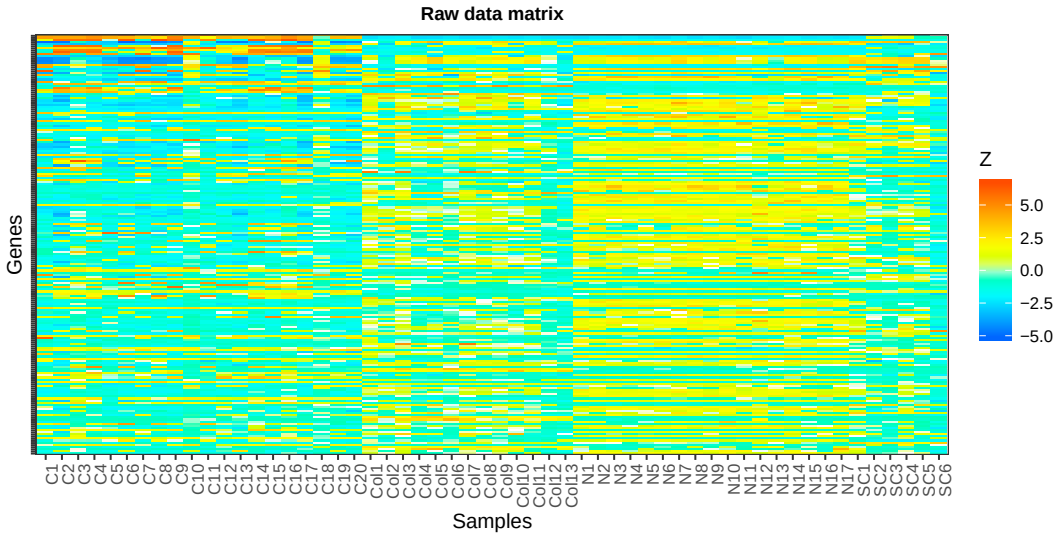

Figure 19: The heatmap of the lung cancer gene expression dataset.

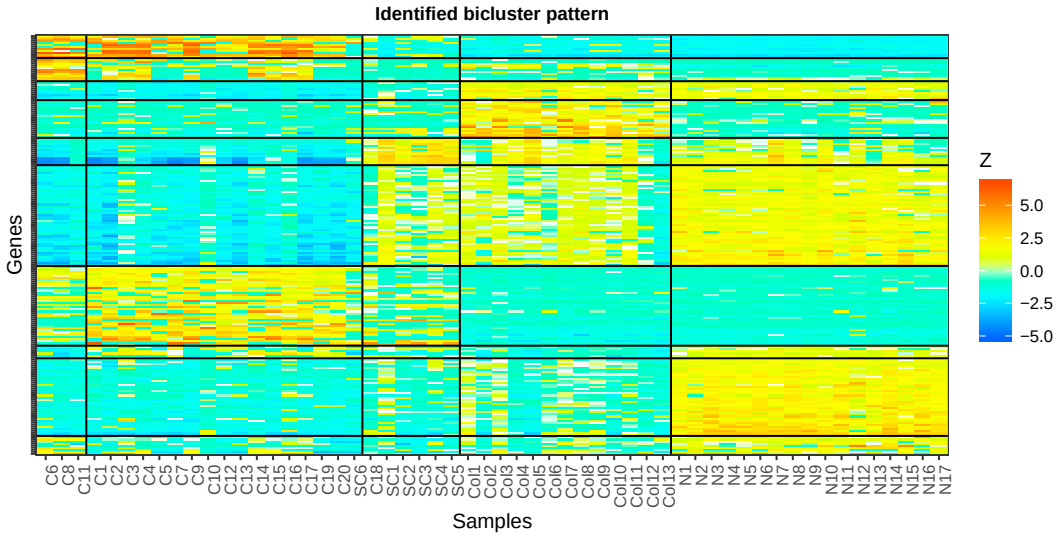

Figure 20: The bicluster pattern identified according to the highest average Rand index criterion in the lung cancer gene expression dataset for $R = 10$, $C = 5$, $\alpha = 10$, and $\beta = 10$.

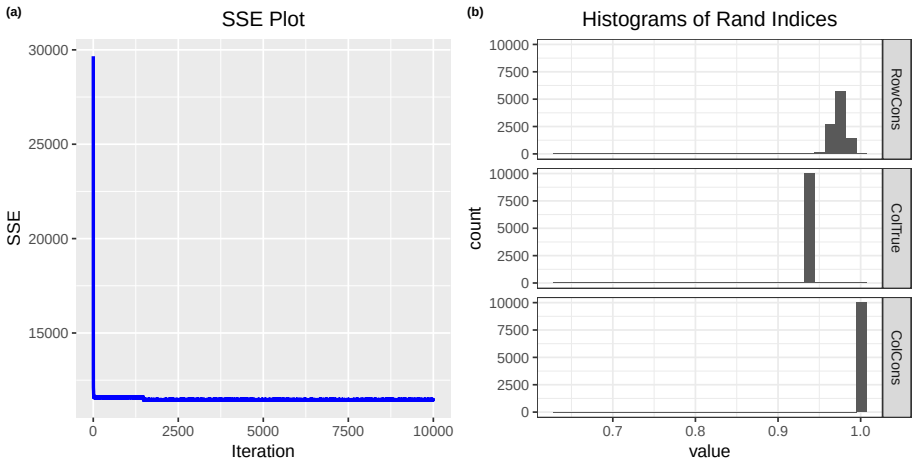

Figure 21: For the biclustering algorithm applied to the lung cancer gene expression dataset with $R = 10$, $C = 5$, $\alpha = 10$, and $\beta = 10$: (a) the total SSE plot, (b) the histograms of Rand indices. Note that, the "true" column clustering is known.

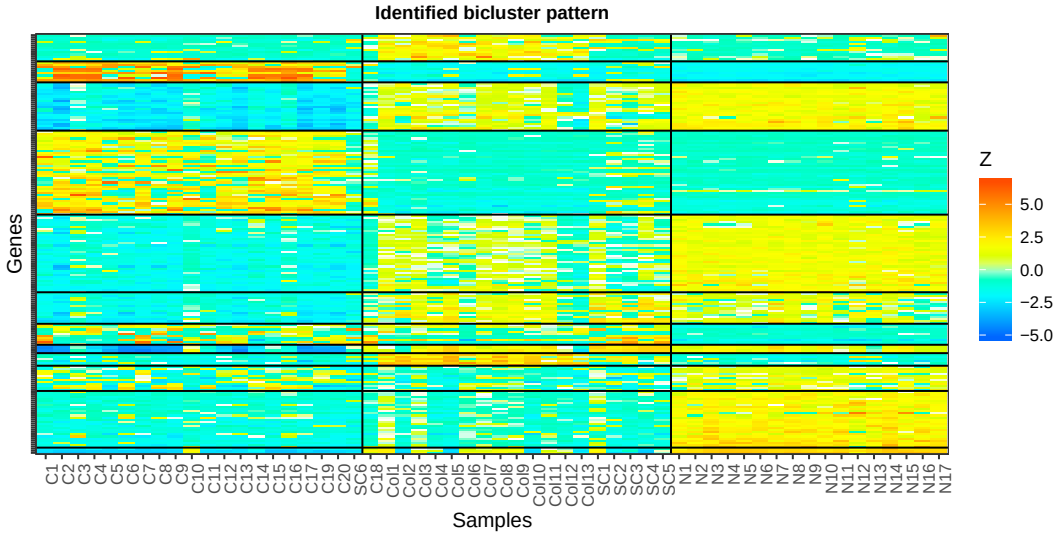

Figure 22: The bicluster pattern identified according to the highest average Rand index criterion in the lung cancer gene expression dataset for $R = 12$, $C = 5$.

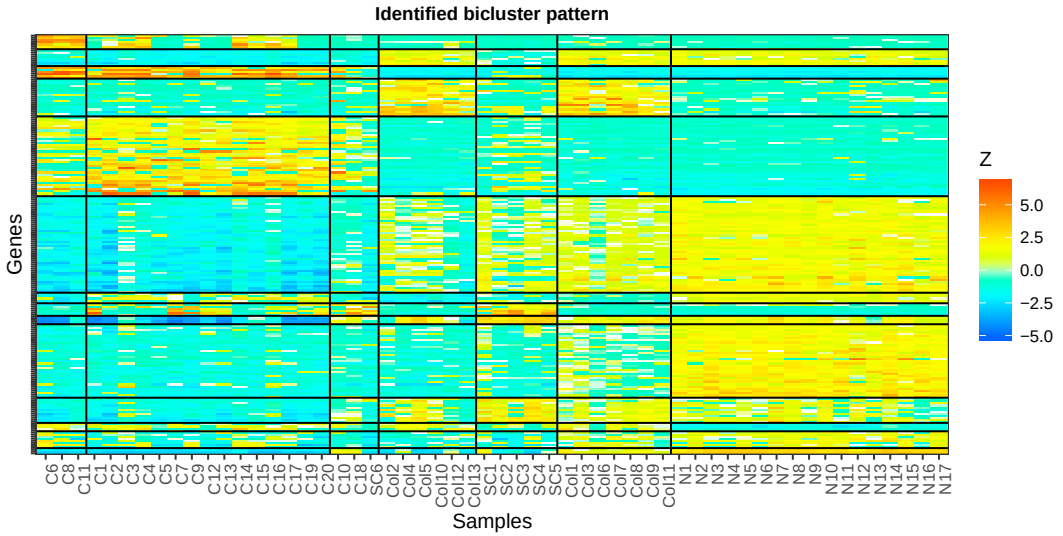

Figure 23: The bicluster pattern identified according to the highest average Rand index criterion in the lung cancer gene expression dataset for $R = 15$, $C = 8$.

C: Application to a real dataset (Section 8.5)

The following plots support the analysis of the agricultural yield dataset.

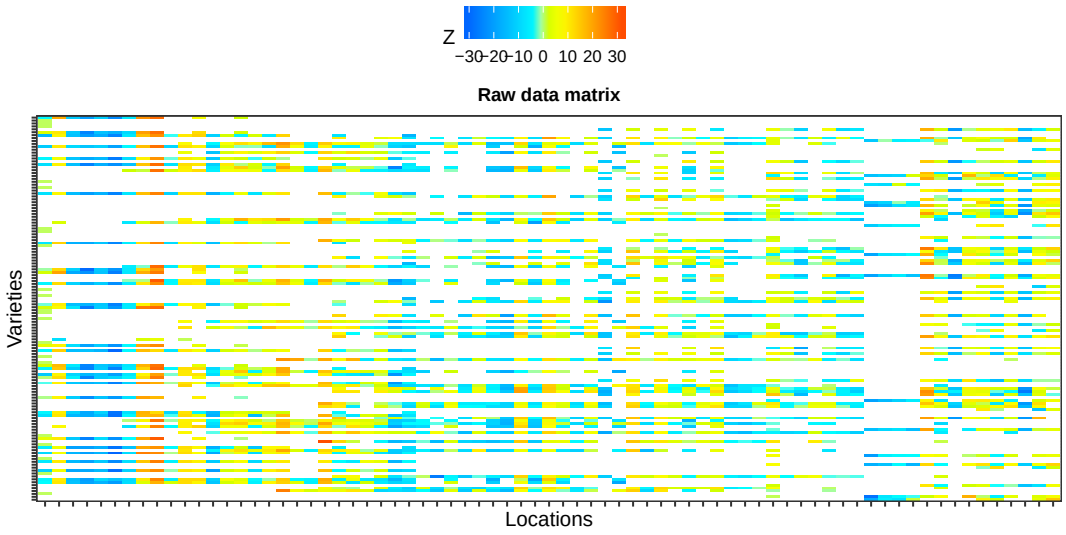

Figure 24: Heatmap of the agricultural yield dataset.

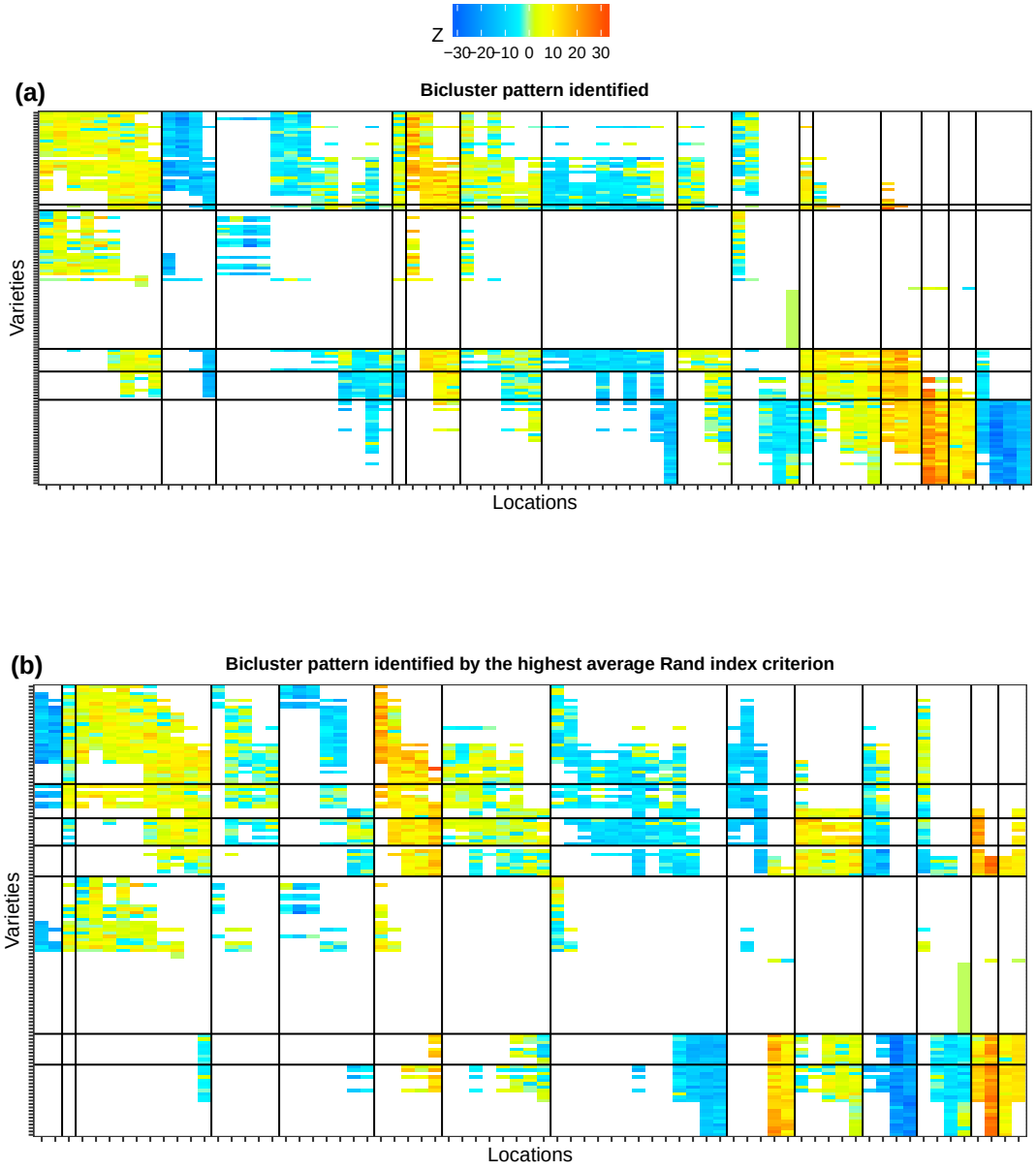

Figure 25: The bicluster patterns identified within the agricultural yield dataset according to the highest average Rand index criterion under the assumptions that (a) the missingness is informative as per the proposed biclustering model, and (b) the missingness is at random.

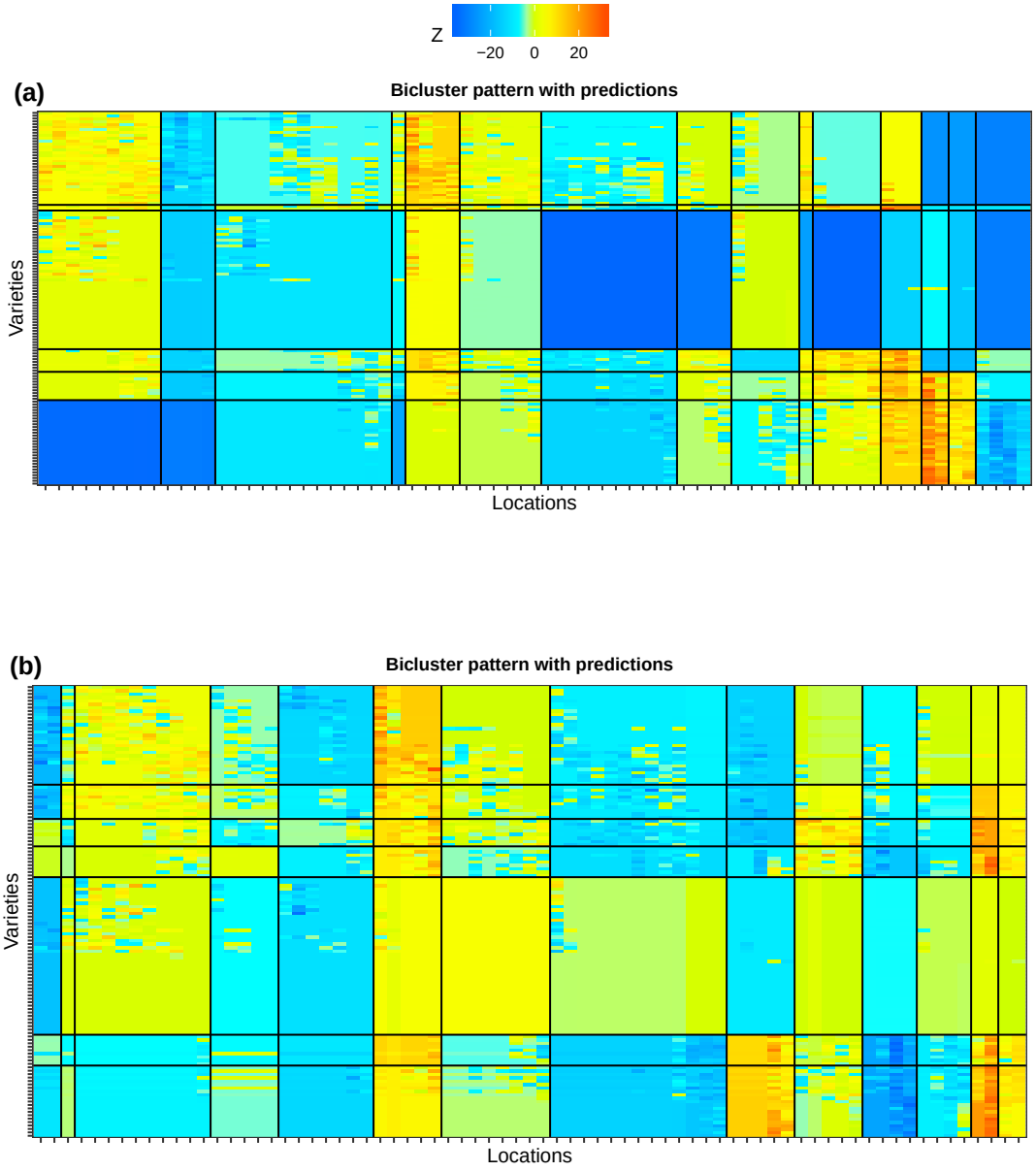

Figure 26: The bicluster patterns identified within the agricultural yield dataset according to the highest average Rand index criterion with the respective predictions under the assumptions that (a) the missingness is informative as per the proposed biclustering model, and (b) the missingness is at random.

Table 5: A summary of total SSE, log posterior, and penalized SSE values computed across $T = 10,000$ complete MCMC iterations with the proposed biclustering model (with assumption of informative missingness) and reported against each distinct combination of numbers of non-empty row and column clusters denoted by p and q respectively. Marked in bold are values of the minimum total SSE, the maximum log posterior, the minimum penalized SSE, their corresponding (p, q) combination, and the (p, q) combination that has the highest average Rand index across all iterates.

p	q	freq	mean SSE	min SSE	max SSE	mean LogPost	min LogPost	max LogPost	min pSSE
6	**13**	281	61287.70	60822.76	69815	-5437.96	-7845.77	-5194.89	**82.784**
6	14	30	60211.18	60211.18	60211.18	-5209.48	-5236.91	-5191.75	88.780
6	**15**	**9688**	59151.02	**59121.80**	59946.52	-4820.16	-5207.62	**−4726.28**	94.772
7	13	1	74283.53	74283.53	74283.53	-13966.06	-13966.06	-13966.06	95.871

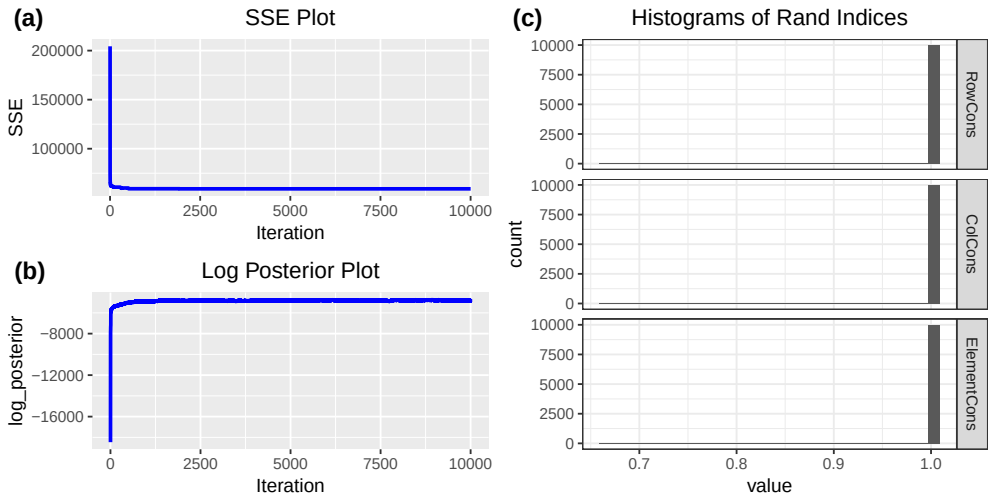

Figure 27: For the biclustering algorithm with the assumption of informative missingness applied to the agricultural yield dataset: (a) the total SSE plot, (b) the log posterior plot, and (c) the histograms of row, column, and element-wise Rand index values computed between two successive iterates along the MCMC chain.

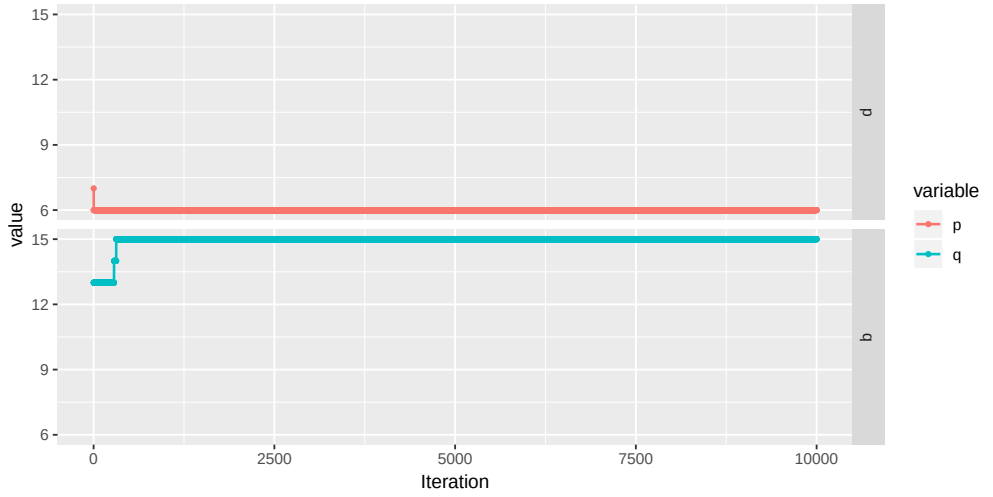

Figure 28: The running plots of the numbers of non-empty row clusters (p) and column clusters (q) against the MCMC iterations for the biclustering algorithm with the assumption of informative missingness applied to the agricultural yield dataset.