Review of iClusterVB: A Fast Integrative Clustering and Feature Selection Approach for High-Dimensional Data

Hadiseh Azadehyaei (T00726889)
Feng Gu (T00751197)

Winter 2025

1 Data source and data description

In this section, we demonstrate the functionality of `iClusterVB` for integrative clustering and feature selection using a simulated multi-omic dataset. This simulation is designed to closely mimic the structure of real-world biological data typically encountered in genomics and precision medicine studies.

The simulated dataset includes data from $N = 240$ individuals, each with measurements across $R = 4$ different data types. Specifically, two of these data types represent **continuous variables**—such as gene or mRNA expression levels—while the third data type consists of **count data**, resembling DNA copy number variations. The fourth data type is **binary**, capturing the presence or absence of mutations or other genomic aberrations. This combination of data types reflects a common setup in integrative genomic analyses, where multiple omic layers are collected for each subject.

We assume that the underlying true structure of the data consists of $K = 4$ distinct clusters, corresponding to four biological subgroups. These clusters are **balanced in size**, meaning each cluster contains 25% of the individuals ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$).

Each data type contains $p_r = 500$ features, for a total of $p = 2000$ features across all data types. Among these, **only 10%** (i.e., 50 features per data type) are informative and contribute to the clustering structure. These informative features are also referred to as **relevant or discriminative features**, as they capture the variations that distinguish between clusters. The remaining 90% of the features in each data type serve as **irrelevant noise features**, which do not contribute to the clustering and are expected to be filtered out during the feature selection process.

The distribution of these useful features across the four clusters is structured, and the table below summarizes how the relevant features are assigned within each cluster and data type.

1.1 Exploratory Data Analysis (EDA)

To identify and visualize the valid normal distributions from the simulated dataset, we performed the following sampling and testing operations:

- **Random Sampling:** We randomly sampled 30 variables from each of the three different distributions (normal, poisson, and multinomial), resulting in a total of 90 variables for testing.
- **Statistical Testing:** For each sampled variable, we conducted the corresponding statistical tests:
  - A normality test (e.g., Shapiro-Wilk test) for variables assumed to follow a normal distribution.
  - A Poisson test for variables assumed to follow a Poisson distribution.
  - A multinomial test for variables assumed to follow a multinomial distribution.
- **Validation and Visualization:** Variables that passed their respective tests were visualized to determine whether they represent truly valid features or are merely low-probability events arising from noise distributions.

Table 1: Distribution of relevant and noise features across clusters in each data view of the simulated data

Data View	Cluster	Distribution
1 (Continuous)	Cluster 1	$\mathcal{N}(10, 1)$ (Relevant)
	Cluster 2	$\mathcal{N}(5, 1)$ (Relevant)
	Cluster 3	$\mathcal{N}(-5, 1)$ (Relevant)
	Cluster 4	$\mathcal{N}(-10, 1)$ (Relevant)
	-	$\mathcal{N}(0, 1)$ (Noise)
2 (Continuous)	Cluster 1	$\mathcal{N}(-10, 1)$ (Relevant)
	Cluster 2	$\mathcal{N}(-5, 1)$ (Relevant)
	Cluster 3	$\mathcal{N}(5, 1)$ (Relevant)
	Cluster 4	$\mathcal{N}(10, 1)$ (Relevant)
	-	$\mathcal{N}(0, 1)$ (Noise)
3 (Binary)	Cluster 1	Bernoulli(0.05) (Relevant)
	Cluster 2	Bernoulli(0.2) (Relevant)
	Cluster 3	Bernoulli(0.4) (Relevant)
	Cluster 4	Bernoulli(0.6) (Relevant)
	-	Bernoulli(0.1) (Noise)
4 (Count)	Cluster 1	Poisson(50) (Relevant)
	Cluster 2	Poisson(35) (Relevant)
	Cluster 3	Poisson(20) (Relevant)
	Cluster 4	Poisson(10) (Relevant)
	-	Poisson(2) (Noise)

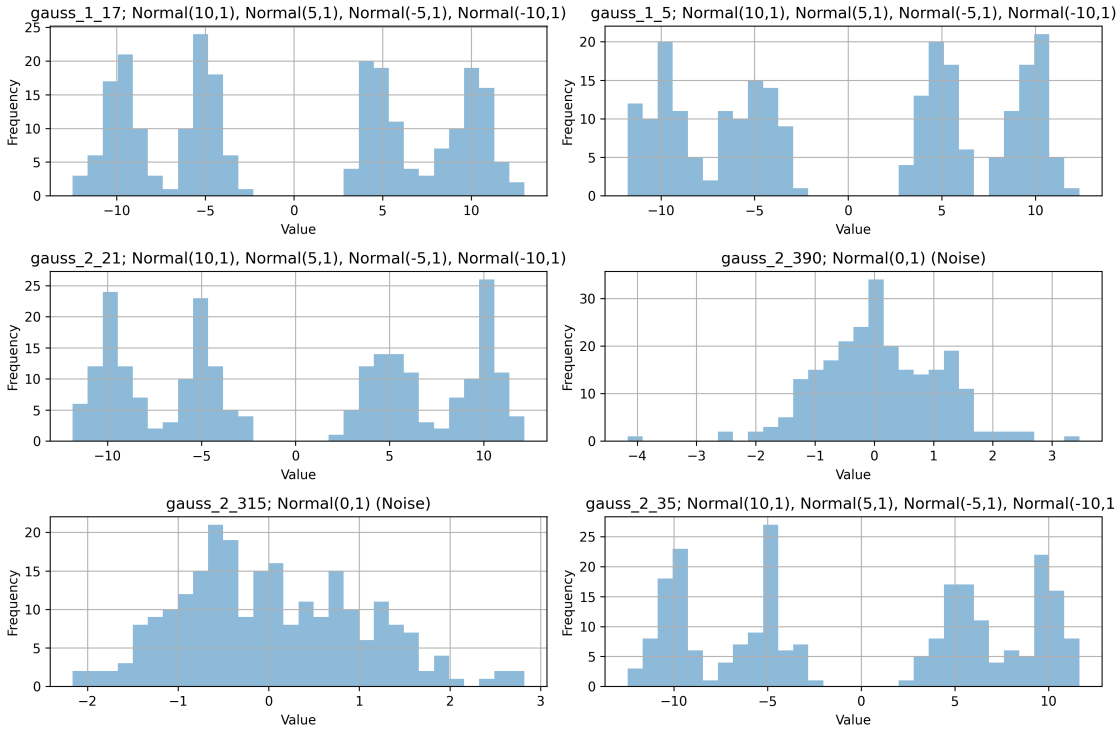

Figure 1: Histograms of normal distributions that pass the normality test(correspond to table 6)

This process helps to filter out irrelevant noise features and retain only the informative variables that contribute to the clustering structure.

Figure 1 displays histograms of variables simulated from normal distributions. Variables with titles in-

cluding "(Noise)" correspond to noise components sampled from $\mathcal{N}(0,1)$, while those without such a label represent relevant features derived from mixtures of $\mathcal{N}(10,1)$, $\mathcal{N}(5,1)$, $\mathcal{N}(-5,1)$, and $\mathcal{N}(-10,1)$. These relevant variables show clear multimodal patterns, indicating the presence of structured information. In contrast, the noise variables exhibit unimodal, symmetric distributions centered around zero, lacking informative structure. The absence of "(Noise)" in the title marks the variables that passed the normality test and were deemed statistically significant for downstream modeling.

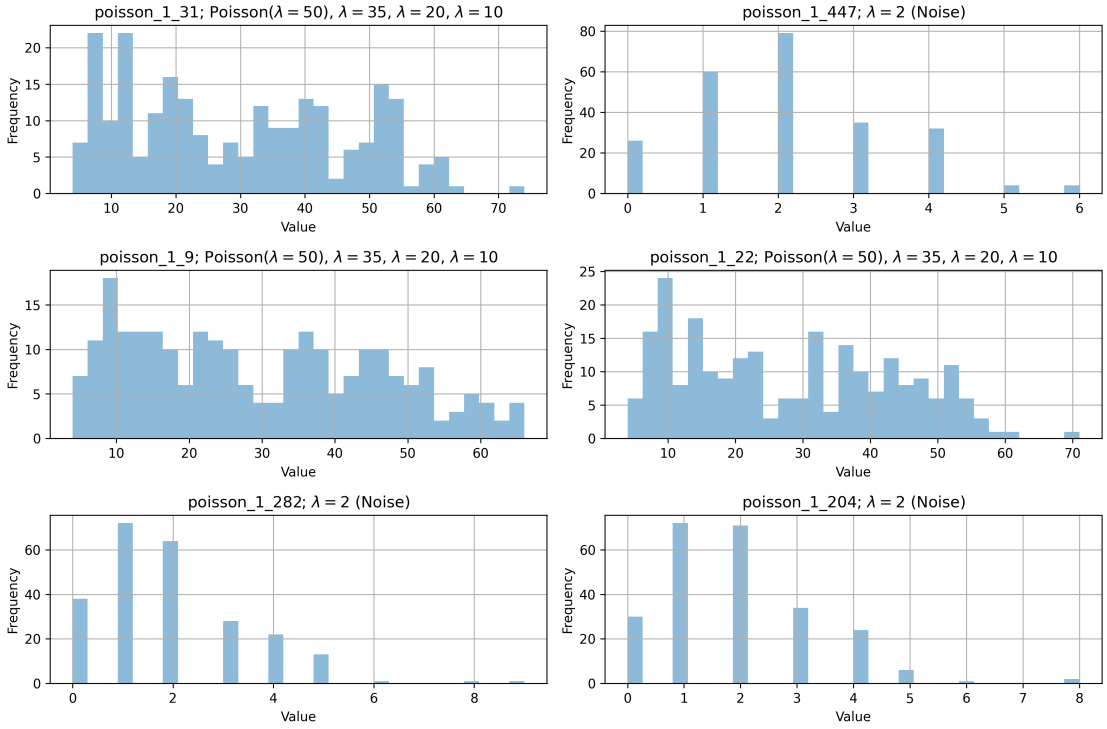

Figure 2: Histograms of normal distributions that pass the poission test (corresponding to table 7)

Figure 2 shows histograms of variables generated from Poisson distributions with different rate parameters λ. Variables labeled with multiple λ values (e.g., $\lambda = 50, 35, 20, 10$) represent informative features composed of mixed Poisson signals. These histograms exhibit broader, multi-modal or skewed distributions, suggesting meaningful variability across groups, which is useful for clustering tasks. In contrast, variables marked as "(Noise)"—such as those with $\lambda = 2$ only—display narrow, concentrated distributions around small integer values. These noise variables do not carry discriminative structure and are thus considered irrelevant for clustering.

1.2 Necessity of testing in EDA for iClusterVB algorithm

The testing operations described above (numerical information are stored in tables 6, 7, 8 in appendix) demonstrate the effectiveness of statistical tests in identifying suitable variables for downstream analysis. By applying normality, Poisson, and multinomial tests, we were able to filter out irrelevant noise features and retain only the informative variables that exhibit meaningful patterns or variability. These selected variables are well-suited for integrative clustering and feature selection tasks, making them ideal candidates for input into the `iClusterVB` algorithm. This ensures that the algorithm operates on high-quality data, enhancing its ability to uncover biologically meaningful clusters and informative features.

2 iClusterVB Implementation on simulation data

In this section, we show how the **iClusterVB** method works using simulated data. The data includes 240 individuals and 4 different types of data views. These data types are often found in genomics studies:

- Two continuous views (like gene expression)
- One binary view (like mutation: yes or no)
- One count view (like DNA copy number)

Each view has 500 features, so in total we have 2000 features. But only 10% of features (50 per view) are useful for clustering. The rest are just noise (not important).

The true number of clusters is 4, and each cluster has the same number of individuals (60 each).

We combined these datasets into a list and specified their types using a vector. Before running the model, we changed the 0 values in the binary data to 2, because the algorithm does not accept zeros. Then, we used the iClusterVB function to run the clustering model with a maximum of 8 clusters. The model correctly found 4 clusters. After that, we used summary to check the results and piplot and chmap to visualize selected features and clusters.

Table 2: Summary of Clustering and Variable Selection Results

Category	**Description**	**Value**
Clustering Hyper-Parameters/Result		
Number of individuals	Total number of observations	240
Maximum clusters specified	User-defined input	6
Number of clusters found	Determined by algorithm	4
Individuals per cluster	Cluster sizes (equal)	60
Variable Selection by View (Threshold: 0.5)		
View 1 (Gaussian)	Variables selected	57 out of 500
View 2 (Gaussian)	Variables selected	58 out of 500
View 3 (Multinomial)	Variables selected	63 out of 500
View 4 (Poisson)	Variables selected	68 out of 500

The clustering algorithm identified a total of 4 clusters from the simulated dataset, despite a user-specified maximum of 6 clusters. Each cluster contains exactly 60 individuals, summing to a total of 240 individuals.

Variable selection results indicate the number of variables exceeding a posterior inclusion probability threshold of 0.5 in each view. Specifically, 57 and 58 variables were selected in View 1 and View 2, respectively (both Gaussian views), while 63 variables were selected in View 3 (Multinomial view), and 68 variables in View 4 (Poisson view), out of a total of 500 variables per view. These results suggest meaningful contributions from all data types, with slightly stronger signals from the Multinomial and Poisson views.

Table 3: Cross-tabulation of Predicted vs. True Cluster Memberships

Predicted Cluster	**True Cluster 1**	**True Cluster 2**	**True Cluster 3**	**True Cluster 4**
1	0	0	60	0
2	0	60	0	0
3	60	0	0	0
6	0	0	0	60

Table 3 presents the cross-tabulation between the predicted cluster labels and the true underlying cluster assignments. Each predicted cluster perfectly matches one of the true clusters without any misclassification.

Specifically, predicted cluster 1 matches true cluster 3, cluster 2 matches true cluster 2, cluster 3 matches true cluster 1, and cluster 6 corresponds exactly to true cluster 4. This result indicates a perfect clustering performance, with all 240 individuals correctly assigned to their respective groups.

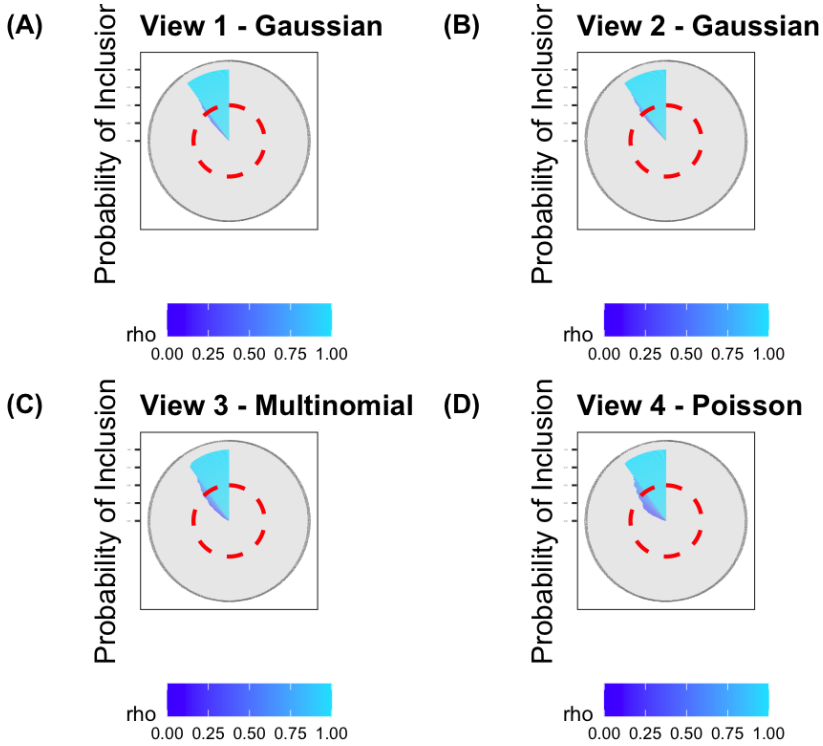

Figure 3: PIP plot showing the posterior inclusion probabilities (PIPs) of features across all data views.

Figure 3 illustrates the posterior probability of variable inclusion across the four different data views: (A) and (B) correspond to the two Gaussian views, (C) corresponds to the Multinomial view, and (D) to the Poisson view. In each subplot, the radial axis represents the probability of inclusion for individual variables, with higher values indicating stronger evidence of variable relevance. The angular separation reflects the clustering structure among variables.

The color gradient encodes the ρ value, which governs the correlation or dependence strength across features. Bright blue regions (high ρ) indicate confidently selected variables, while darker tones indicate low inclusion probability or noise. Across all views, the most informative variables are highlighted clearly with high inclusion probabilities concentrated in narrow angular bands, consistent with the true signal structure of the simulated dataset.

Figure 4 shows heatmaps of the four data views used for clustering. Clear block patterns in Views 1 and 2 (Gaussian) indicate that these features distinguish clusters well. View 3 (Multinomial) also displays cluster-specific structure in the upper rows. While View 4 (Poisson) is mostly sparse, the top block shows variation across clusters, suggesting some informative features. Together, these heatmaps validate that each view contributes to identifying distinct clusters.

3 iClusterVB Implementation on Breast Cancer data

In this section, we transition from working with simulation data to applying our methodology to a more realistic dataset.

Specifically, we utilize the Breast Cancer dataset obtained from Kaggle. This dataset provides a practical context to evaluate the performance and applicability of the iClusterVB implementation in addressing real-world challenges.

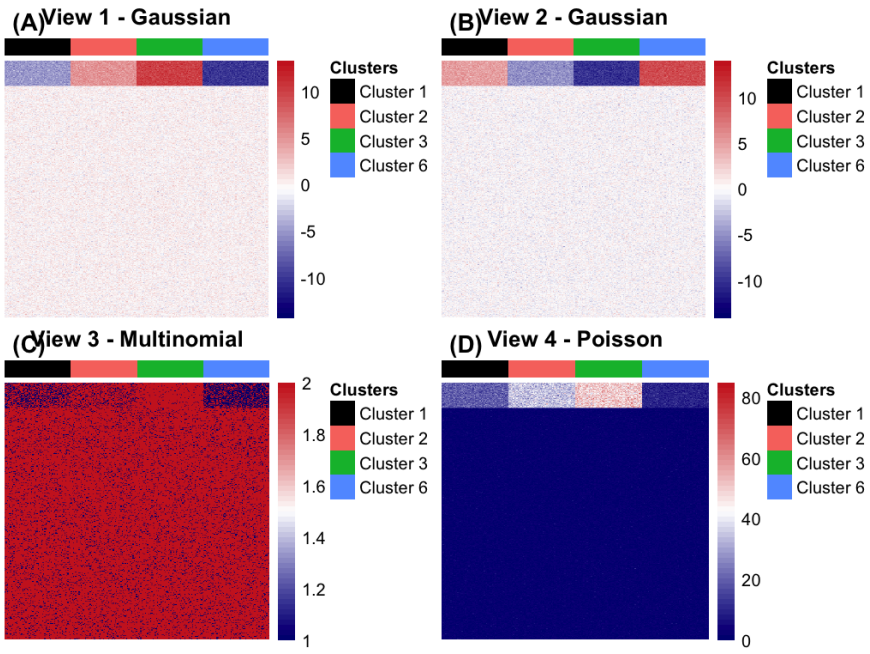

Figure 4: Cluster heatmap showing the clustering structure and variable selection across all data views.

There are no missing values in the dataset, and all features are numeric(except the target variable(M & B)), which provide us convenient conditions for applying the iClusterVB method.

Before taking statistical testing, we did some data preprocessing. We encoded the target variable and split the data based on the target variable, got two subsets of data, one for M and one for B. After checking the distributions of the two data sets, we found there were heavy skewness and outliers in the data, therefore we make a simple logarithm transformation on the data [1].

3.1 Distribution Testing

By normality testing, we found that most variables are not normally distributed, only four variables are normally distributed, which are: *area_se, texture_worst, concave points_worst, symmetry_worst.*

Hence, we will take these four variables as valid normally distributed features as one of the inputs for the iClusterVB method.

We also did Poisson testing on the data, and we found that most variables are not Poisson distributed, only one variable *texture_se* is Poisson distributed.

Other variables that are not shown in the table are not suitable for the Poission test, and we assumed they are not Poisson distributed.

To multinomial(bernoulli) test, we found that all variables are not interget-valued, hence they are not suitable for multinomial test. Therefore, we will not take multinomial test in this case.

3.2 iClusterVB Results

Clustering Summary on Breast Cancer Subset

The clustering algorithm was applied to a real subset of the breast cancer dataset, yielding the following results:

Table 5 presents the clustering summary for a subset of the breast cancer dataset. The algorithm successfully identified 6 clusters from 357 individuals, matching the user-specified maximum number of

[1] In later practice, other transfomration can be used here, taking log-transfomration is only for convenience of demonstration here

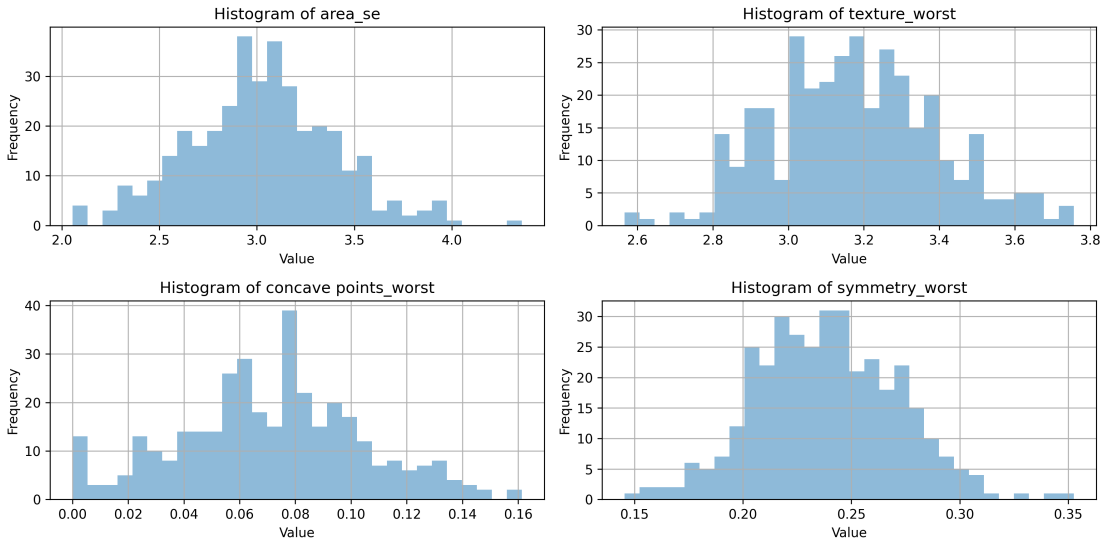

Figure 5: Histograms of normal distributions that pass the normality test(corresponding to table 9)

Table 4: Chi-square test for all variables in the breast cancer dataset

index	Variable	Chi2 Statistic	Poisson-p-value
3	texture_mean	25.2086	0.0000
4	perimeter_mean	334.8591	0.0000
5	area_mean	232.2412	0.0000
13	texture_se	1.5309	0.2160
14	perimeter_se	63.7165	0.0000
15	area_se	108.8869	0.0000
22	radius_worst	235.0098	0.0000
23	texture_worst	137.9469	0.0000
24	perimeter_worst	353.2600	0.0000
25	area_worst	354.8223	0.0000

Table 5: Clustering Summary for Breast Cancer Subset

Metric	Result
Total number of individuals	357
Maximum number of clusters (user input)	6
Number of clusters determined	6
Variables selected (posterior inclusion probability ¿ 0.5)	
View 1 - Gaussian	2 out of 4
View 2 - Poisson	2 out of 2

Cluster Membership Distribution

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
57	55	80	46	98	21

clusters. Cluster sizes ranged from 21 to 98. Variable selection results show that 2 out of 4 Gaussian variables and all 2 Poisson variables were selected based on a posterior inclusion probability threshold of 0.5,

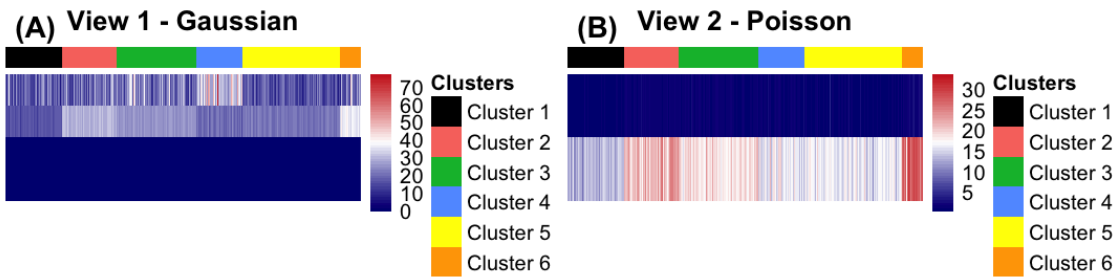

Figure 6: Heatmap of the clustering results on breast cancer data

indicating their importance in distinguishing clusters across the two views.

Figure 6 shows the clustering heatmaps for the breast cancer dataset based on two data views: (A) Gaussian and (B) Poisson. Each column corresponds to an individual, and rows represent variables. The color bars at the top denote the cluster membership of each individual. In both views, clear block patterns are observed, indicating distinct variable activity across different clusters. This separation suggests that the clustering algorithm effectively identified subgroups with different underlying characteristics in both Gaussian and Poisson features.

4 Appendix

4.1 Tables

Table 6: Normality test for Gaussian distribution (first 1000 columns in the simulation data)

Index	Variable	Skewness	Kurtosis	Normality p-value
0	gauss_2_396	-0.1079	-0.0766	0.7817
1	gauss_1_225	0.0277	-0.1564	0.9313
2	gauss_2_18	-0.0052	-1.5656	0.0000
3	gauss_1_90	-0.1858	0.3446	0.2363
4	gauss_1_375	0.0521	-0.0305	0.9374
5	gauss_2_217	-0.0205	-0.1222	0.9708
6	gauss_1_490	0.0701	-0.0306	0.8950
7	gauss_1_260	-0.0380	0.0563	0.8946
8	gauss_1_124	0.1376	-0.2938	0.4495
9	gauss_1_400	-0.1274	-0.3238	0.4176
10	gauss_1_167	0.1571	-0.3698	0.2736
11	gauss_1_18	-0.0073	-1.5921	0.0000
12	gauss_1_318	0.0031	0.1366	0.8125
13	gauss_1_220	0.1701	-0.4797	0.1043
14	gauss_1_308	-0.0471	-0.3887	0.3883
15	gauss_1_385	0.0960	-0.1883	0.7422
16	gauss_2_210	0.0282	0.3164	0.5181
17	gauss_2_354	-0.0808	0.2698	0.5219
18	gauss_1_369	0.0930	0.0298	0.7920
19	gauss_1_147	-0.2302	0.1885	0.2441
20	gauss_1_50	-0.0317	-1.6111	0.0000
21	gauss_1_405	-0.0403	0.0832	0.8590
22	gauss_1_117	0.0661	-0.3219	0.5403
23	gauss_2_179	-0.0018	-0.3278	0.5762
24	gauss_2_249	0.3049	0.6479	0.0258
25	gauss_1_68	-0.1215	-0.1024	0.7268
26	gauss_1_299	0.2415	-0.1434	0.2882
27	gauss_1_118	-0.2193	0.0210	0.3537
28	gauss_2_48	0.0079	-1.5732	0.0000
29	gauss_2_484	0.1617	-0.2456	0.4565

Table 7: Poisson test for Poisson distribution (1500 to 2000 columns in the simulation data)

Index	Variable	Statistic	Poisson p-value
0	poisson_1_347	8.1418	0.4197
1	poisson_1_130	3.3996	0.8457
2	poisson_1_278	2.0546	0.9568
3	poisson_1_31	1896309619.6291	0.0000
4	poisson_1_233	2.1521	0.9509
5	poisson_1_405	7.6532	0.3642
6	poisson_1_343	4.8117	0.6829
7	poisson_1_458	4.6286	0.5923
8	poisson_1_336	9.0090	0.3415
9	poisson_1_222	7.1677	0.4116
10	poisson_1_440	6.7327	0.3463
11	poisson_1_357	7.3086	0.3975
12	poisson_1_341	9.9632	0.1907
13	poisson_1_113	5.7099	0.4565
14	poisson_1_420	1.8409	0.9337
15	poisson_1_447	12.7249	0.0476
16	poisson_1_248	7.0096	0.5356
17	poisson_1_417	7.0637	0.2159
18	poisson_1_194	2.7987	0.8337
19	poisson_1_9	6684971.6159	0.0000
20	poisson_1_61	2.9649	0.8132
21	poisson_1_22	952499618.9309	0.0000
22	poisson_1_407	0.9116	0.9887
23	poisson_1_282	41.5326	0.0000
24	poisson_1_129	8.1200	0.4218
25	poisson_1_421	7.0042	0.4284
26	poisson_1_204	24.9668	0.0008
27	poisson_1_427	3.0758	0.8779
28	poisson_1_258	10.4206	0.1080
29	poisson_1_383	5.0726	0.6511

Table 8: Bernoulli test for Bernoulli distribution (1000 to 1500 columns in the simulation data)

Index	**Variable**	**Chi2 Statistic**	**p-value**	**Estimated** p	**# 0s**	**# 1s**
0	multinomial_1_353	0.7407	0.3894	0.1167	212	28
1	multinomial_1_345	1.6667	0.1967	0.0750	222	18
2	multinomial_1_196	1.6667	0.1967	0.0750	222	18
3	multinomial_1_103	1.1574	0.2820	0.1208	211	29
4	multinomial_1_83	1.1574	0.2820	0.0792	221	19
5	multinomial_1_499	1.6667	0.1967	0.0750	222	18
6	multinomial_1_22	106.6667	0.0000	0.3000	168	72
7	multinomial_1_138	0.4167	0.5186	0.0875	219	21
8	multinomial_1_360	4.6296	0.0314	0.1417	206	34
9	multinomial_1_72	0.7407	0.3894	0.1167	212	28
10	multinomial_1_486	0.0000	1.0000	0.1000	216	24
11	multinomial_1_101	2.9630	0.0852	0.0667	224	16
12	multinomial_1_90	0.0463	0.8296	0.0958	217	23
13	multinomial_1_42	161.1574	0.0000	0.3458	157	83
14	multinomial_1_271	0.0463	0.8296	0.0958	217	23
15	multinomial_1_252	0.0000	1.0000	0.1000	216	24
16	multinomial_1_110	0.7407	0.3894	0.1167	212	28
17	multinomial_1_415	0.7407	0.3894	0.1167	212	28
18	multinomial_1_122	0.7407	0.3894	0.0833	220	20
19	multinomial_1_230	0.4167	0.5186	0.0875	219	21
20	multinomial_1_403	1.6667	0.1967	0.0750	222	18
21	multinomial_1_88	0.4167	0.5186	0.1125	213	27
22	multinomial_1_126	7.8241	0.0052	0.1542	203	37
23	multinomial_1_130	1.1574	0.2820	0.0792	221	19
24	multinomial_1_439	0.4167	0.5186	0.1125	213	27
25	multinomial_1_16	89.6296	0.0000	0.2833	172	68
26	multinomial_1_340	3.7500	0.0528	0.1375	207	33
27	multinomial_1_293	0.0000	1.0000	0.1000	216	24
28	multinomial_1_64	0.0000	1.0000	0.1000	216	24
29	multinomial_1_82	0.4167	0.5186	0.0875	219	21

Table 9: Normality test for all variables in the breast cancer dataset

Index	**Variable**	**Skewness**	**Kurtosis**	**Normality p-value**
0	id	6.9517	48.7647	0.0000
1	radius_mean	-0.4930	0.2500	0.0006
2	texture_mean	0.3480	0.1690	0.0207
3	perimeter_mean	-0.5277	0.3695	0.0002
4	area_mean	-0.5650	0.3883	0.0001
5	smoothness_mean	0.6002	1.5932	0.0000
6	compactness_mean	1.0849	1.7566	0.0000
7	concavity_mean	2.9089	15.0911	0.0000
8	concave points_mean	0.8717	0.8859	0.0000
9	symmetry_mean	0.5683	1.0837	0.0000
10	fractal_dimension_mean	1.6090	4.2304	0.0000
11	radius_se	1.1020	2.0694	0.0000
12	texture_se	0.5976	0.5768	0.0000
13	perimeter_se	0.5015	-0.0299	0.0009
14	area_se	0.1348	0.2530	0.3185
15	smoothness_se	1.4897	2.9455	0.0000
16	compactness_se	2.1325	5.4932	0.0000
17	concavity_se	5.4665	45.6831	0.0000
18	concave points_se	2.0728	9.9365	0.0000
19	symmetry_se	1.3380	3.0752	0.0000
20	fractal_dimension_se	4.2709	26.7974	0.0000
21	radius_worst	-0.4131	0.0001	0.0073
22	texture_worst	0.1375	-0.2272	0.3988
23	perimeter_worst	-0.4131	0.0143	0.0072
24	area_worst	-0.4709	0.0885	0.0017
25	smoothness_worst	0.2805	0.2833	0.0477
26	compactness_worst	0.8333	0.8487	0.0000
27	concavity_worst	1.6051	5.2030	0.0000
28	concave points_worst	0.0221	-0.2058	0.7610
29	symmetry_worst	0.1265	0.1762	0.4333
30	fractal_dimension_worst	1.3515	2.8535	0.0000