STAT 5244 – Unsupervised Learning

Homework 2

Name: Chuyang Su UNI: cs4570

1 Mixture Models

1.1 EM Algorithm Derivation

Since we model count-valued data, assume each observation $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{N}_0^p$ is generated from a finite mixture of *independent* Poisson distributions:

$$p(x_i;\, \pi, \lambda) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \frac{e^{-\lambda_{kj}} \lambda_{kj}^{x_{ij}}}{x_{ij}!}, \quad \pi_k \geq 0,\ \sum_{k=1}^{K} \pi_k = 1,\ \lambda_{kj} > 0.$$

Here $\pi = (\pi_1, \ldots, \pi_K)$ are mixture weights and $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kp})$ are component-wise Poisson means.

Latent variables. Introduce latent indicators $z_{ik} \in \{0, 1\}$ with $\sum_{k=1}^{K} z_{ik} = 1$, where $z_{ik} = 1$ if x_i comes from component k. The complete-data likelihood is

$$L_c(\pi, \lambda) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{p} \frac{e^{-\lambda_{kj}} \lambda_{kj}^{x_{ij}}}{x_{ij}!} \right]^{z_{ik}}.$$

Taking logs and dropping constants independent of (π, λ) (i.e., $\log x_{ij}!$) gives the complete-data log-likelihood

$$\ell_c(\pi, \lambda) \propto \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k + \sum_{j=1}^{p} \left(x_{ij} \log \lambda_{kj} - \lambda_{kj} \right) \right].$$

E-step Derivation. Since the latent indicators z_{ik} are unobserved, we take their conditional expectation under the current parameters. Define

$$\gamma_{ik} := \mathbb{E}[z_{ik} \mid x_i; \pi^{(t)}, \lambda^{(t)}] = P(z_{ik} = 1 \mid x_i; \pi^{(t)}, \lambda^{(t)}),$$

which represents the posterior probability that observation x_i belongs to component k.

Using Bayes' theorem,

$$P(z_{ik} = 1 \mid x_i; \pi^{(t)}, \lambda^{(t)}) = \frac{P(z_{ik} = 1; \pi^{(t)})\, P(x_i \mid z_{ik} = 1; \lambda^{(t)})}{P(x_i; \pi^{(t)}, \lambda^{(t)})}.$$

Each term can be expressed as:

$$P(z_{ik} = 1; \pi^{(t)}) = \pi_k^{(t)}, \quad P(x_i \mid z_{ik} = 1; \lambda^{(t)}) = \prod_{j=1}^{p} \frac{e^{-\lambda_{kj}^{(t)}} (\lambda_{kj}^{(t)})^{x_{ij}}}{x_{ij}!},$$

and

$$P(x_i; \pi^{(t)}, \lambda^{(t)}) = \sum_{\ell=1}^{K} \pi_\ell^{(t)} \prod_{j=1}^{p} \frac{e^{-\lambda_{\ell j}^{(t)}} (\lambda_{\ell j}^{(t)})^{x_{ij}}}{x_{ij}!}.$$

Substituting these expressions into Bayes' rule yields:

$$\gamma_{ik} = \frac{\pi_k^{(t)} \prod_{j=1}^{p} e^{-\lambda_{kj}^{(t)}} (\lambda_{kj}^{(t)})^{x_{ij}} / x_{ij}!}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} \prod_{j=1}^{p} e^{-\lambda_{\ell j}^{(t)}} (\lambda_{\ell j}^{(t)})^{x_{ij}} / x_{ij}!}.$$

Since the term $\prod_{j=1}^{p} x_{ij}!$ does not depend on k, it cancels out between numerator and denominator. Therefore, the final expression for the responsibilities is:

$$\boxed{\gamma_{ik} = \frac{\pi_k^{(t)} \prod_{j=1}^{p} e^{-\lambda_{kj}^{(t)}} (\lambda_{kj}^{(t)})^{x_{ij}}}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} \prod_{j=1}^{p} e^{-\lambda_{\ell j}^{(t)}} (\lambda_{\ell j}^{(t)})^{x_{ij}}}}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K.$$

M-step. We maximize

$$Q(\pi, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[\log \pi_k + \sum_{j=1}^{p} \left(x_{ij} \log \lambda_{kj} - \lambda_{kj} \right) \right]$$

subject to $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, and $\lambda_{kj} > 0$.

Update for π_k. Introduce a Lagrange multiplier η for the simplex constraint:

$$\mathcal{L}(\pi, \eta) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \gamma_{ik} \right) \log \pi_k + \eta \left(1 - \sum_{k=1}^{K} \pi_k \right).$$

Setting the partial derivatives to zero,

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{\sum_{i=1}^{n} \gamma_{ik}}{\pi_k} - \eta = 0 \quad \Longrightarrow \quad \pi_k = \frac{\sum_{i=1}^{n} \gamma_{ik}}{\eta}.$$

Summing over k and using $\sum_{k=1}^{K} \pi_k = 1$ gives

$$1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\eta} \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_{ik} = \frac{1}{\eta} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} = \frac{1}{\eta} \sum_{i=1}^{n} 1 = \frac{n}{\eta} \implies \eta = n.$$

Hence

$$\boxed{\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}}.$$

(Concavity: $\partial^2 \mathcal{L} / \partial \pi_k^2 = -(\sum_i \gamma_{ik}) / \pi_k^2 < 0$.)

Update for λ_{kj}. For each (k, j), the terms of Q that involve λ_{kj} are

$$Q_{kj}(\lambda_{kj}) = \sum_{i=1}^{n} \gamma_{ik}\big(x_{ij} \log \lambda_{kj} - \lambda_{kj}\big).$$

Differentiate and set to zero:

$$\frac{\partial Q_{kj}}{\partial \lambda_{kj}} = \sum_{i=1}^{n} \gamma_{ik} \left(\frac{x_{ij}}{\lambda_{kj}} - 1 \right) = 0 \quad \Longrightarrow \quad \lambda_{kj} = \frac{\sum_{i=1}^{n} \gamma_{ik} x_{ij}}{\sum_{i=1}^{n} \gamma_{ik}}.$$

(Concavity: $\partial^2 Q_{kj}/\partial \lambda_{kj}^2 = -\sum_i \gamma_{ik} x_{ij}/\lambda_{kj}^2 < 0$ when $\lambda_{kj} > 0$.) Therefore

$$\boxed{\lambda_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik} x_{ij}}{\sum_{i=1}^{n} \gamma_{ik}}}.$$

1.2 Interpretation and Comparison of Poisson and Gaussian Mixture Models.

Both the Poisson Mixture Model (PMM) and the Gaussian Mixture Model (GMM) successfully converged and achieved almost identical clustering performance on the author–chapter word-frequency data.

The PMM converged in 16 iterations with a clustering purity of **0.8989**, while the GMM converged in 14 iterations with a purity of **0.9001**. The learned author–cluster mappings were consistent across models: Cluster 0 corresponds to *Shakespeare*, Cluster 1 to *London*, and Clusters 2–3 to *Austen*, indicating that Austen's writing exhibits two stylistically distinct sub-modes.

Examining the estimated centroids revealed interpretable linguistic patterns. The Shakespeare cluster assigns high weights to archaic forms such as *thou*, *thy*, and *hath*, capturing the syntax of Early Modern English. London's centroid emphasizes neutral verbs like *was*, *had*, and *were*, representing narrative realism, while Austen's two clusters differ primarily in pronoun and modal-verb usage: one dominated by *she, her, would, could* (social dialogue tone), and the other by *my, our, only, such* (introspective narration).

By inspecting the soft responsibilities γ_{ik}, chapters with $\max_k \gamma_{ik} < 0.6$ were identified as *low-certainty chapters*. These typically occur near stylistic transitions or among authors with overlapping vocabularies. Such ambiguous sections highlight that soft clustering provides richer insights than hard assignments.

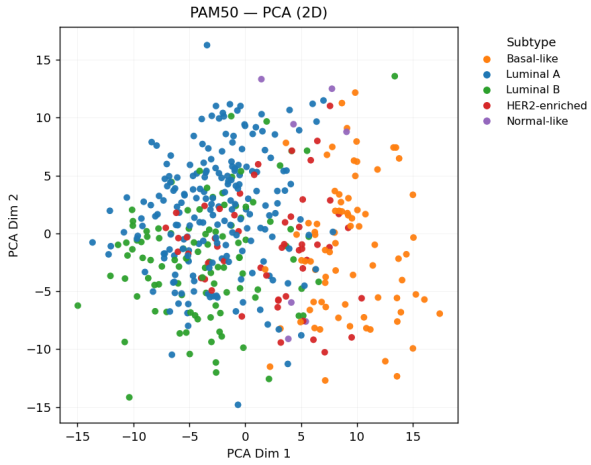
Overall, both mixture models achieve roughly 90% purity and uncover meaningful stylistic structures. While the GMM yields a marginally higher purity, the Poisson mixture remains theoretically more appropriate for discrete count data and offers comparable empirical performance.

2 Open-Ended Cluster Analysis - Breast Cancer gene expression data.

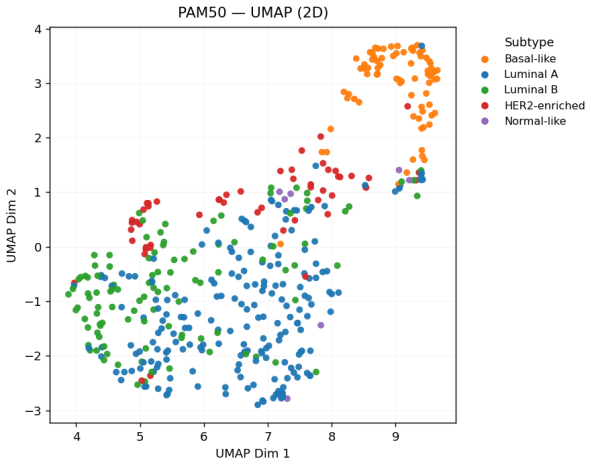
2.1 Apply clustering techniques to explore this data set.

To explore the internal structure of the BRCA gene-expression dataset, I first applied a two-step dimensionality reduction approach using **PCA** followed by **UMAP**.

Based on conclusions drawn in Homework 1, this strategy provides an effective balance between global and local structure preservation: PCA removes redundant noise variables while retaining the main variance directions, and UMAP further compresses the PCA-transformed space, producing a low-dimensional embedding that reveals local neighborhood structure and facilitates visual interpretation. Following the empirical results from Homework 1, I selected 30 principal components (the "elbow" point in the cumulative variance curve) for PCA, and adopted UMAP parameters $n\_neighbors = 10$, $min\_dist = 0.0$, and $n\_components = 2$ or 10, where the two-dimensional embedding was used for visualization and the ten-dimensional representation served as the feature space for clustering. The results are shown below:



(a) PAM50 subtypes visualized in PCA(2D) space.



(b) PAM50 subtypes visualized in UMAP(2D) space.

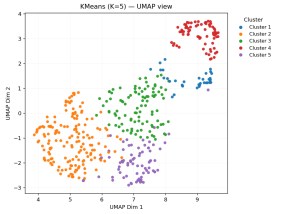
Subsequent clustering was performed using seven algorithms: **KMeans**, **Gaussian Mixture Model (GMM)**, **Spectral Clustering**, **DBSCAN**, and **Agglomerative Clustering** with three linkages (Ward, Average, and Complete). Given that the biological subtype reference (PAM50) contains five classes, all parametric clustering methods were set to $K = 5$, while DBSCAN automatically determined the number of clusters.

Among these, **Spectral Clustering** was treated as a special case. Because it internally constructs a similarity graph and performs Laplacian eigen-decomposition—conceptually similar to UMAP's graph-based embedding—it is inappropriate to apply Spectral Clustering directly in UMAP space. Doing so would perform a second spectral decomposition on an already nonlinear manifold, likely distorting the intrinsic geometry. Therefore, Spectral Clustering was conducted on the PCA(30D) features instead, ensuring consistency with its

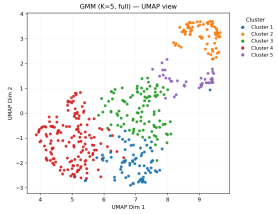
theoretical formulation.

For the Agglomerative Clustering, I compared only the **Ward**, **Average**, and **Complete** linkages, excluding **Single linkage** for three reasons: (1) it is highly sensitive to outliers and prone to chaining effects, which are amplified in high-dimensional continuous data; (2) its computational cost is high while its silhouette scores are unstable; and (3) the remaining three linkages already represent the major clustering behaviors.

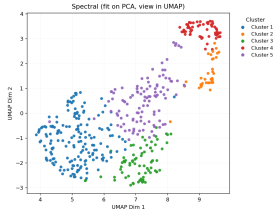
All the clustering results expect of hierachical clusterings is shown below:



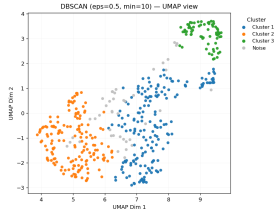
(a) K-Means



(b) Gaussian Mixture Model

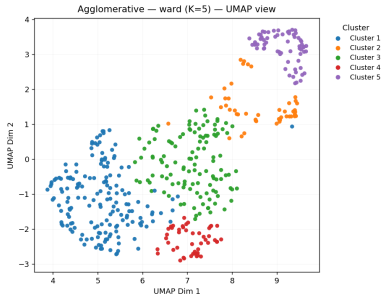


(c) Spectral Clustering in PCA(30)

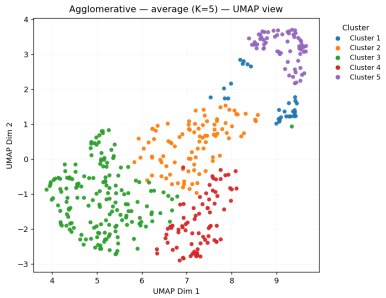


(d) DBSCAN

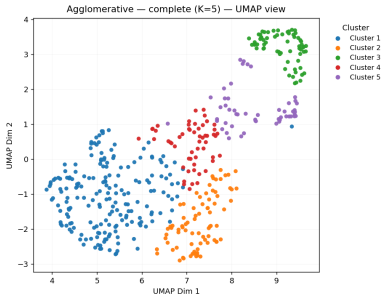
And three types of hierachical clusterings' results are:



(a) Ward



(b) Average



(c) Complete

Insights from the clustering results. Rather than directly comparing algorithmic performance, the focus here is on what each clustering method reveals about the *data manifold itself*. Across methods, two key structural patterns emerge. First, the dataset contains both compact and diffuse regions: **Basal-like** and **Normal-like** samples consistently form dense, isolated clusters, while **Luminal A** and **Luminal B** exhibit gradual overlap, suggesting a continuous transition rather than distinct boundaries. Second, the overall data distribution is non-spherical and heterogeneous in density, which explains why algorithms assuming isotropic clusters (KMeans, GMM, Ward) perform similarly and fail to separate overlapping subtypes.

DBSCAN highlights these density variations most clearly—it detects three main groups and labels the remaining sparse regions as noise. Although this underestimates the total number of subtypes, it accurately reflects the intrinsic density imbalance of the dataset: the rarest subtypes (**HER2-enriched**, **Normal-like**) are naturally absorbed as low-density regions. **Average linkage** produces smoother transitions that capture the gradual relationship between Luminal subtypes, while **Complete linkage** emphasizes inter-cluster separation. Taken together, these clustering outcomes indicate that the BRCA expression data contain

both discrete and continuous subtype structures—Basal-like being compact and distinct, whereas Luminal A/B lie on a continuum of transcriptional profiles—consistent with known biological heterogeneity among breast cancer subtypes.

2.2 Implement cluster validation techniques.

In this section, I evaluate the clustering results obtained in 2(a) through systematic parameter tuning and validation. For algorithms that require specifying the number of clusters K (*KMeans*, *GMM*, *Spectral Clustering*, and *Agglomerative Clustering* with three linkages), I performed a grid search over $K = 2, \ldots, 9$ and computed three complementary validation criteria:

- **Silhouette Score** — measures intra-cluster cohesion and inter-cluster separation;
- **Stability** — estimated via bootstrap resampling, quantified by the median Adjusted Rand Index (ARI) between full-data and bootstrap partitions;
- **Generalizability** — assessed by a train–test split, where cluster assignments were propagated from training to testing samples (via centroids or kNN label transfer), and the Silhouette score on the test set was reported.

For **DBSCAN**, which is density-based and does not rely on K, a two-dimensional grid search was conducted across $\varepsilon \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and min_samples $\in \{5, 10, 15\}$. Each configuration was evaluated using the same three validation criteria. Additionally, the k-distance graph was inspected to locate the elbow point corresponding to a suitable ε threshold.

All validation computations were performed on the same reduced feature spaces, PCA (30D) for Spectral Clustering, UMAP(10D) for others, to ensure consistency with the embeddings generated in Section 2(a).

Results of Validation Metrics

Table 1 summarizes the best-performing configurations for each clustering method according to the combined Silhouette, Stability, and Generalizability criteria.

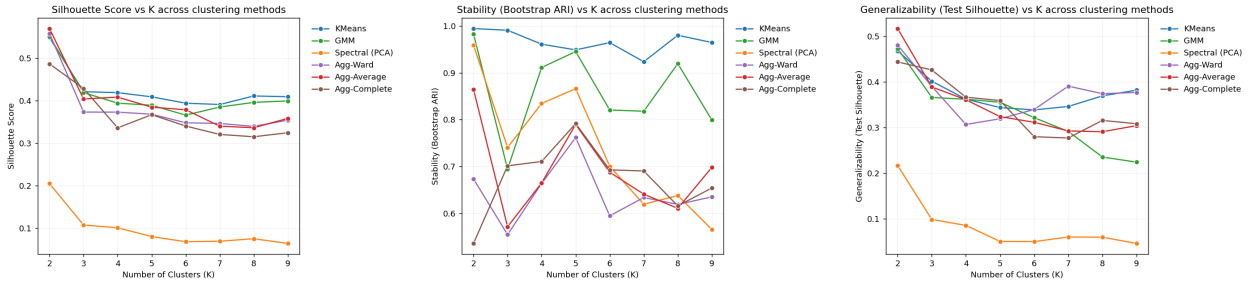
Table 1: Best validation results per method.

Method	K	ε	min_samples	Silhouette	Stability	Generalizability
KMeans	2	–	–	0.549	**0.994**	0.468
GMM	2	–	–	0.551	0.983	0.473
Spectral (PCA)	2	–	–	0.205	0.959	0.217
Agglomerative (Ward)	2	–	–	0.558	0.673	0.481
Agglomerative (Average)	2	–	–	**0.570**	0.864	**0.517**
Agglomerative (Complete)	2	–	–	0.487	0.535	0.444
DBSCAN	–	0.3	15	0.292	0.929	0.344

Across all K-based algorithms, the optimal number of clusters was consistently $K = 2$. This indicates that the BRCA gene-expression manifold is dominated by two large-scale density regions rather than five fully separated clusters, which is biologically plausible since Luminal A/B and HER2-enriched subtypes form a continuous spectrum, while Basal-like and Normal-like remain distinct.

Discussion and Selection of the Best Result

Among all evaluated methods, the **Agglomerative Clustering (Average linkage, $K = 2$)** achieved the highest Silhouette (0.57) together with strong stability (0.86) and generalizability (0.52), providing the best trade-off between cluster compactness and robustness. KMeans and GMM yielded similar two-cluster partitions with excellent stability (> 0.98), indicating consistent global structure but limited ability to separate overlapping subtypes. Spectral Clustering performed worse due to over-smoothing in PCA space, and DBSCAN emphasized density variation rather than discrete boundaries, identifying a few compact regions while labeling sparse points as noise.



(a) Silhouette vs K (b) Stability vs K (c) Generalizability vs K

Figure 4: Validation metrics across different cluster numbers.

Figure 4 illustrates the variation of validation metrics with respect to the number of clusters K.

2.3 Interpret your cluster findings with respect to the metadata provided.

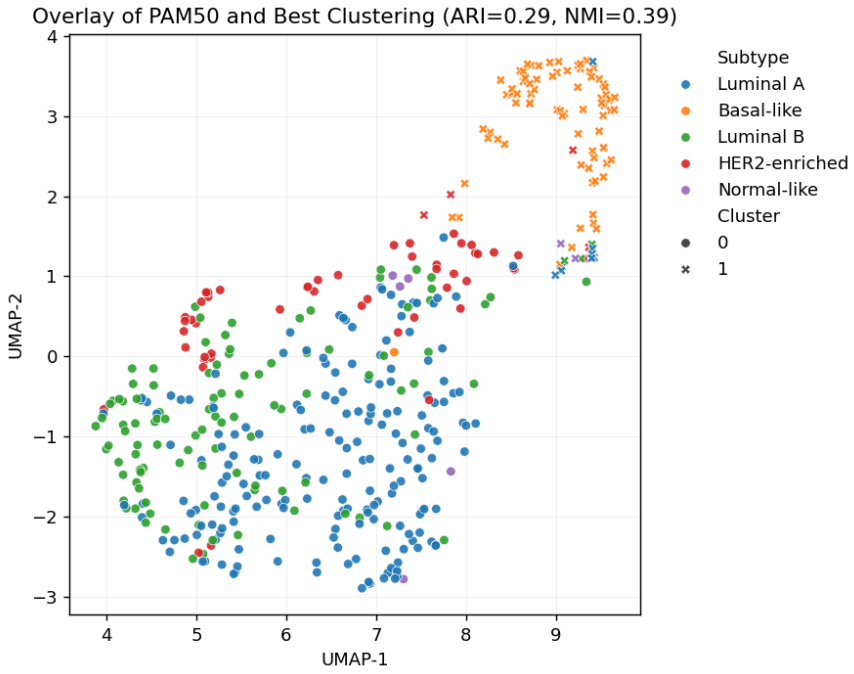


Figure 5: Overlay of PAM50 subtypes and the best clustering result (Agglomerative Average, $K = 2$) in UMAP(2D) space.

The two clusters discovered by the optimal Agglomerative (Average) model ($K = 2$) show a clear correspondence with the clinical PAM50 categories. As shown in Figure 5, one cluster primarily groups **Basal-like** and **Normal-like** tumors, while the other aggregates **Luminal A**, **Luminal B**, and **HER2-enriched** samples. This separation captures the well-known *luminal–basal dichotomy* in breast cancer biology: luminal tumors are typically hormone-receptor positive and exhibit similar expression patterns, whereas basal tumors represent the triple-negative subtype with distinct molecular characteristics. Therefore, the unsupervised clustering not only aligns with the known clinical taxonomy, but also highlights that the BRCA gene-expression landscape is dominated by two broad transcriptional regimes rather than five sharply separated classes.

A Appendix: Code Implementation