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 $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{N}_0^p$ 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 $\pi = (\pi_1, \dots, \pi_K)$ are mixture weights and $\lambda_k = (\lambda_{k1}, \dots, \lambda_{kp})$ 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

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

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

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

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:

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

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.

A Appendix: Code Implementation