A Report on "Combining Mixture Components for Clustering" by Baudry et al.

Branden Olson

1 Background and motivation

Many situations arise in which collected data points must be clustered into similar subgroups. For example, we might wish to cluster DNA sequences from a number of species into phylogenetic families, or categorize X-ray data readings by the galaxies from which the waves originated. In general, we consider clusters to be contiguous, densely-populated regions of a feature space, separated by contiguous, relatively empty regions. The field of clustering contains a broad and vast set of methods based on various assumptions of the underlying data-generating mechanism.

One such category, known as model-based clustering, assumes that the data points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ were generated from a mixture model and can be assigned cluster labels based on features of the distribution. For continuous data, model-based clustering typically assumes that each point $\mathbf{X}_i$ is generated from a multivariate Gaussian mixture model. That is, given K mixture components, we decompose the density of $\mathbf{x}_i$ as

$$f(\mathbf{x}_i|K, \boldsymbol{\theta}_K) = \sum_{i=k}^{K} p_k \varphi\left(\mathbf{x}_i; \mu_k, \Sigma_k\right).$$

Here, $p_k, k = 1, \ldots, K$ are the mixture proportions, or the marginal probabilities for each component to generate a new data point a priori, so that $p_k \geq 0 \, \forall k$ and $\sum_{k=1}^{K} p_k = 1$; $\varphi(\cdot; \mu_k, \Sigma_k)$ is the d-variate Gaussian density with mean μ_k and covariance matrix Σ_k; and $\boldsymbol{\theta}_K = (\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (p_1, \ldots, p_{K-1}, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_K)$.

It's useful to frame the problem as observing incomplete or censored data from a "complete" experiment where the components that generate each data point are known. Notationally, each "complete" data point is a pair $(\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{z}_i \in \mathbb{R}^K$ such that $z_{ik} = \mathbb{1}\left(\mathbf{x}_i \text{ was generated from component } k\right)$. If K is fixed, parameters can be estimated using EM algorithm by incorporating the density of each complete data pair,

$$f(\mathbf{x}_i, \mathbf{z}_i|K, \boldsymbol{\theta}_K) = \prod_{k=1}^{K} \left[p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right]^{z_{ik}}$$

whose derivation is given in Appendix A. The EM algorithm is appropriate here since the true component assignments $\mathbf{z}_i$ can be considered as "missing" data on which we can compute conditional expectations given the observed data.

The goal in many clustering problems is to choose the number of components K either based on previous knowledge, or perhaps without any other information. Dasgupta and Raftery (1998) were the first to propose using the BIC to select the number of components, which in this context is synonymous with the number of clusters. That is, given a set of candidates $\{K_{\min}, \ldots, K_{\max}\}$, compute the MLE estimates for each K, and then choose the K which minimizes the BIC (Bayesian information criterion), defined as

$$\mathrm{BIC}(K) = \log p(\mathbf{x}|K, \widehat{\boldsymbol{\theta}}_{K,\mathrm{MLE}}) - \frac{\nu_K \log(n)}{2}.$$

Here, $f(\mathbf{x}|K, \boldsymbol{\theta}_K) = \prod_{i=1}^{n} f(\mathbf{x}_i|K, \boldsymbol{\theta}_K)$ is the density of the full dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and ν_K is the number of parameters of the K-component model. The BIC criterion can be viewed as an approximation to the integrated likelihood

$$\mathrm{IL}(\mathbf{x}|K) = \int_{\Theta_K} f(\mathbf{x}|K, \boldsymbol{\theta}_K) \mathrm{d}\pi(\boldsymbol{\theta}_K)$$

where $\pi(\cdot)$ is a prior distribution over the parameter space $\mathbf{\Theta}_K$. Under certain regularity conditions, the BIC can be shown to generally estimate the number of components well. However, a drawback of this approach is that it implicitly assumes that each mixture component corresponds to its own cluster which is not always desriable. In other words, some clusters would be better estimated as mixtures of Gaussians themselves, which is unattainable with this approach.

Biernacki et al (2001) modify this approach by introducing the ICL criterion, namely,

$$\widehat{\text{ICL}} \approx \log f(\mathbf{x}, \widehat{\mathbf{z}}|K, \widehat{\theta}_{K,\text{MLE}}) + \frac{\nu_K \log(n)}{2}$$

an approximation of the integrated complete likelihood of the data. Here, $f(\mathbf{x}, \mathbf{z}|K, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i|K, \boldsymbol{\theta})$ is the complete density of the full dataset. Since we don't know the values of $\mathbf{z}_i$, we must plug in estimates $\widehat{\mathbf{z}}_i$ using, for example, maximum a posteriori (MAP) estimates. Analogously to the BIC, this criterion approximates the integrated complete likelihood

$$\text{ICL} = \int_{\Theta_K} f(\mathbf{x}, \mathbf{z}|K, \boldsymbol{\theta}_K)\, \mathrm{d}\pi(\boldsymbol{\theta}_K).$$

It turns out that the ICL can also be seen as a penalization of the BIC criterion based on the mean Shannon entropy of the fitted model with K components,

$$\text{Ent}(K) = -\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\widehat{\boldsymbol{\theta}}_K) \log t_{ik}(\widehat{\boldsymbol{\theta}}_K)$$

The mathematical details can be found in Appendix A.

Appendix A: Theoretical foundations for ICL

The likelihood for one observation is

$$f(\mathbf{x}_i|K, \theta_K) = \sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)$$

Thus, assuming independence, the observed likelihood is

$$f(\mathbf{x}|K, \theta_K) = \prod_{i=1}^{n}\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)$$

That is, the observed log-likelihood is

$$\begin{aligned} L(\theta_K|\mathbf{x}, K) &= \log\left(\prod_{i=1}^{n}\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) \\ &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) \end{aligned}$$

Now, noting that for a unit vector $\mathbf{z}_i \in \mathbb{R}^K$ with kth element 1,

$$\mathsf{Pr}\left(\mathbf{z}_i\right) = \mathsf{Pr}\left(z_{ik} = 1\right) = p_k$$

the complete likelihood for one observation is

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{z}_i) &= f(\mathbf{x}_i|\mathbf{z}_i)\mathsf{Pr}\left(\mathbf{z}_i\right) \\ &= \sum_{k=1}^{K} \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\mathbb{1}\left(\mathbf{z}_{ik} = 1\right) p_k \\ &= \prod_{k=1}^{K} \left[p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right]^{z_{ik}} \end{aligned}$$

so that the complete likelihood is

$$f(\mathbf{x}, \mathbf{z}|K, \theta_K) = \prod_{i=1}^{n}\prod_{k=1}^{K} [p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)]^{z_{ik}}$$

yielding the complete log-likelihood

$$\mathrm{CL}(\theta_K|\mathbf{x}, \mathbf{z}, K) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log\left(p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right).$$

Let t_{ik} be the conditional probability that $\mathbf{x}_i$ comes from component k. Then from Bayes's theorem,

$$\begin{aligned}
t_{ik} &= t_{ik}(K, \boldsymbol{\theta}_K) \\
&= \mathsf{Pr}\left(\text{component } k \mid \mathbf{x}_i, K, \boldsymbol{\theta}_K\right) \\
&= \frac{\mathsf{Pr}\left(\text{component } k, \mathbf{x}_i \mid K, \boldsymbol{\theta}_K\right)}{f(\mathbf{x}_i|K, \boldsymbol{\theta}_K)} \\
&= \frac{\mathsf{Pr}\left(\mathbf{x}_i|\text{component } k, K, \boldsymbol{\theta}_K\right)\mathsf{Pr}\left(\text{component } k|K, \boldsymbol{\theta}_K\right)}{\sum_{j=1}^{K} p_j \varphi(\mathbf{x}_i|\mu_j, \Sigma_j)} \\
&= \frac{\varphi(\mathbf{x}_i|\mu_k, \Sigma_k) \cdot p_k}{\sum_{j=1}^{K} p_j \varphi(\mathbf{x}_i|\mu_j, \Sigma_j)}
\end{aligned}$$

If we define

$$\mathrm{EC}(K|\mathbf{z}) := -\sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik} \log t_{ik}$$

We see that

$$\begin{aligned}
L(\theta_K) - \mathrm{EC}(K) &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) - \left(-\sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik} \log(t_{ik})\right) \\
&= \sum_{i=1}^{n}\left\{\log\left(\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) + \sum_{k=1}^{K} z_{ik} \log\left(\frac{p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} p_j \varphi(\mathbf{x}_i|\mu_j, \Sigma_j)}\right)\right\} \\
&= \sum_{i=1}^{n}\left\{\log\left(\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) + \sum_{k=1}^{K} z_{ik}\left[\log\left(p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) - \log\left(\sum_{j=1}^{K} p_j \varphi(\mathbf{x}_i|\mu_j, \Sigma_j)\right)\right]\right\} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log(p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)) + \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)\right) \underbrace{\left[1 - \sum_{i=1}^{K} z_{ik}\right]}_{=0} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log(p_k \varphi(\mathbf{x}_i|\mu_k, \Sigma_k)) \\
&\equiv \mathrm{CL}(K)
\end{aligned}$$

Seeing $\mathrm{EC}(K|\mathbf{Z})$ as a random variable, we have for a fixed K,

$$\begin{aligned}
\mathsf{E}\left[\mathrm{EC}(K|\mathbf{Z})\right] &= -\sum_{k=1}^{K}\sum_{i=1}^{n} \mathsf{E}\left[Z_{ik}\right] \log t_{ik} \\
&= -\sum_{k=1}^{K}\sum_{i=1}^{n} \mathsf{Pr}\left(Z_{ik} = 1\right) \log t_{ik} \\
&= -\sum_{k=1}^{K}\sum_{i=1}^{n} \mathsf{Pr}\left(X_i \text{ comes from } k\text{th component}\right) \log t_{ik} \\
&\equiv -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik} \log t_{ik} \\
&= \mathrm{Ent}(K),
\end{aligned}$$

the entropy of the matrix $\mathbf{T} = \{t_{ik}\}$.

Now, the integrated likelihood (aka evidence or model evidence) is

$$\begin{aligned}\mathrm{IL}(\mathbf{x}|K,\theta_K) &= \int_{\Theta_K} f(\mathbf{x}|K,\theta_K)\pi(\theta_K|K)\,\mathrm{d}\theta_K \\ &= \int_{\Theta_K} \prod_{i=1}^{n}\sum_{k=1}^{K} p_k\varphi(\mathbf{x}_i|\mu_k,\Sigma_k)\pi(\theta_K|K)\,\mathrm{d}\theta_K\end{aligned}$$

An approximation is given via

$$\begin{aligned}\log \mathrm{IL}(\mathbf{x}|K,\theta_K) &\approx \log f(\mathbf{x}|K,\widehat{\theta}_{K,\mathrm{MLE}}) - \frac{\nu_K \log(n)}{2} \\ &\equiv \mathrm{BIC}(K)\end{aligned}$$

where $\widehat{\theta}_{MLE} = \arg\max_\theta f(\mathbf{x}|K,\theta)$. Analogously, we define the integrated complete likelihood as

$$\begin{aligned}\mathrm{ICL}(\mathbf{x},\mathbf{z}|K) &= \int_{\Theta_K} f(\mathbf{x},\mathbf{z}|K,\theta_K)\pi(\theta_K|K)\,\mathrm{d}\theta_K \\ &= \int_{\Theta_K} \prod_{i=1}^{n}\prod_{k=1}^{K} [p_k\varphi(\mathbf{x}_i|\mu_k,\Sigma_k)]^{z_{ik}}\,\pi(\theta_K|K)\,\mathrm{d}\theta_K\end{aligned}$$

which takes into account evidence for an effective clustering structure. We can introduce a similar approximation to BIC for missing data:

$$\log \mathrm{ICL} \approx \log f(\mathbf{x},\widehat{\mathbf{z}}|K,\widehat{\theta}_{K,\mathrm{MLE}}) + \frac{\nu_K \log(n)}{2}$$

where

$$\widehat{z}_{ik} = \mathbb{1}\left(\arg\max_{1\le\ell\le K} t_{i\ell}(\widehat{\theta}_{MLE}) = k\right).$$

Appendix B: Foundations of Baudry et al. methodology

Suppose we merge clusters k and k' to form $k \cup k'$. Then

$$\begin{aligned}t^K_{i,k\cup k'} &= \mathsf{Pr}\left(\mathbf{x}_i \text{ comes from } (k\cup k')\text{th component}\right) \\ &= \mathsf{Pr}\left(\mathbf{x}_i \text{ comes from } k\text{th or } k'\text{th component}\right) \\ &= \mathsf{Pr}\left(\mathbf{x}_i \text{ comes from } k\text{th component }\right) + \mathsf{Pr}\left(\mathbf{x}_i \text{ comes from } k'\text{th component}\right) \\ &= t^K_{ik} + t^K_{ik'}\end{aligned}$$

Then the resultant entropy (after possibly relabeling the indices) is

$$\begin{aligned}\mathrm{Ent}(K-1) &= -\sum_{k=1}^{K-1}\sum_{i=1}^{n} t_{ik}\log t_{ik} \\ &= -\sum_{i=1}^{n}\left(\sum_{j\neq k\cup k'} t_{ij}\log(t_{ij}) + t_{i,k\cup k'}\log(t_{i,k\cup k'})\right) \\ &= -\sum_{i=1}^{n}\left(\sum_{j\neq k\cup k'} t_{ij}\log(t_{ij}) + (t^K_{i,k}+t^K_{i,k'})\log(t^K_{ik}+t^K_{ik'})\right)\end{aligned}$$

Thus, we have the corresponding difference in entropy:

$$\begin{aligned}
\Delta(K) &\equiv \text{Ent}(K) - \text{Ent}(K-1) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}\log t_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{K-1} t_{ij}\log t_{ij} \\
&= -\sum_{i=1}^{n}\left\{\sum_{j\neq k,k'} t_{ij}\log t_{ij} + t_{ik}\log t_{ik} + t_{ik'}\log t_{ik'} + \sum_{j\neq k\cup k'} t_{ij}\log t_{ij} + t_{i,k\cup k'}\log t_{i,k\cup k'}\right\} \\
&= -\sum_{i=1}^{n}\{t_{ik}\log t_{ik} + t_{ik'}\log t_{ik'}\} + \sum_{i=1}^{n} t_{i,k\cup k'}\log t_{i,k\cup k'}
\end{aligned}$$

Thus, we choose clusters k and k' such that combining them maximizes $\Delta(K)$, i.e., yields the highest decrease in entropy moving from K to $K-1$ clusters, which is good since low entropy means a well-partitioned model

Appendix C: Expressions for ν_K

Recall the BIC criterion, defined as

$$\text{BIC}(K) = \log p(\mathbf{x}|K,\widehat{\theta_K}) - \frac{\nu_K}{2}\log(n).$$

If we assume general covariance matrices Σ_k, then the number of model parameters is

$$\begin{aligned}
\nu_K &= \dim \mathcal{M}_K \\
&= \dim \theta_K \\
&= \dim (p_1,\ldots,p_{K-1},\mathbf{a}_1,\ldots,\mathbf{a}_K) \\
&= \dim (p_1,\ldots,p_K) + \dim (\mathbf{a}_1,\ldots,\mathbf{a}_K) \\
&= K-1+K\dim(\mathbf{a}_1) \\
&= K-1+K\left(\dim(\mu_1,\Sigma_1)\right) \\
&= K-1+K\left[m+m(m+1)/2\right] \\
&= (K+1)\left[m+\frac{m(m+1)}{2}\right]-1.
\end{aligned}$$

Appendix D: Miscellaneous calculations

Conditional model probabilities given the data:

$$\begin{aligned}
\mathsf{Pr}(M_\ell|\mathbf{x}) &= \frac{f(\mathbf{x}|M_\ell)\mathsf{Pr}(M_\ell)}{f(\mathbf{x})} \\
&= \frac{f(\mathbf{x}|M_\ell)\mathsf{Pr}(M_\ell)}{\sum_{r=1}^{m} f(\mathbf{x}|M_r)\mathsf{Pr}(M_r)}
\end{aligned}$$

If $M_1 = \ldots = M_m$, then

$$\mathsf{Pr}(M_\ell|\mathbf{x}) = \frac{f(\mathbf{x}|M_\ell)}{\sum_{r=1}^{m} f(\mathbf{x}|M_r)}.$$