Update 01/04/2020

Stephen Coleman

26/03/2020

General model

For data $X = (x_1, \ldots, x_N)$, where each item $x_i = (x_{i1}, \ldots, x_{iP})$, we use a K-component mixture-model paramaterised by θ to describe the data:

$$p(x_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k f(x|\theta_k).$$

Here $\pi = (\pi_1, \ldots, \pi_K)$ is the proportion of items assigned to each component and θ_k is the component specific parameters.

We assume that there is a common probability density function, $f(\cdot)$, associated with each component (e.g. Gaussian). Independence is assumed between the P features, thus:

$$p(x_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} f(x|\theta_{kp}),$$

where θ_{kp} is the parameters for the p^{th} feature within the k^{th} component (e.g. if we are using a *Gaussian mixture model*, then $\theta_{kp} = (\mu_{kp}, \sigma^2_{kp})$, the mean and standard deviation of the items in the k^{th} component within the p^{th} feature).

In the language of Law, Jain, and Figueiredo (2003), we assume that a subset of the features are *irrelevant*. By this we mean that for a given item x_i,

$$f(x_i|\theta_{kp}) = f(x_i|\theta_{lp}) = g(x_i|\lambda_p) \ \forall \ k, l \in \{1, \ldots, K\}.$$

Thus an irrelevant feature does not contribute any component specific information and is irrelevant to uncovering structure within the data. Let $\Phi = (\phi_1, \ldots, \phi_P)$ be a binary variable indicating the relevance of a feature (i.e. $\phi_p = 1$ if the p^{th} feature is relevant and 0 otherwise). Then our model can be written:

$$p(x_i|\theta, \pi, \Phi) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} f(x_i|\theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}. \tag{1}$$

If we use an allocation variable $z = (z_1, \ldots, z_N)$ to indicate which component the items belong to, we may write:

$$p(x_i|z_i = k, \theta, \pi, \Phi) = \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}$$
$$p(x|z, \theta, \pi, \Phi) = \prod_{i=1}^{N} \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}$$

Simulations

In our simulations we are interested in testing how *consensus inference* compares to Bayesian inference of mixture models in various circumstances. In each simulation we will assume a generative model that can be described by a finite mixture of Gaussian models.

$$p(x_i|z_i = k, \theta, \pi, \Phi) = \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} f(x_i|\theta_p)^{(1-\phi_p)}. \tag{2}$$

where $f(\cdot)$ describes the Gaussian pdf and thus $\theta = (\mu, \sigma^2)$. We assume that the irrelevant variables will be Gaussian in form and that the observed values will be drawn from the same Gaussian distribution for the entire population. Let $P_n = \sum_{p=1}^{P} \phi_p$ be the number of irrelevant features present, and $P_s = P - P_n$ be the number of relevant features present. Then in each simulation we will change various variables associated with this model:

- N: the number of items being clustered;
- P_s: the number of *relevant* features present;
- P_n: the number of *irrelevant* features present;
- K: the true number of subpopulations present;
- π: the proportion of points sampled from each component;
- Δ_μ: the difference between the means associated with each component in each feature;
- σ^2: the standard deviation within each feature for each component; and
- α: the concentration of the Dirichlet distribution π might be generated from (only relevant when π is sampled as explained below).

I would expect that there is some function of the number of samples, the number of informative features, the number of clusters, the distance between component means and the value of σ^2 used that explains how easy it is to resolve the clutering structure. If C' is the true clustering and C^* is that predicted by the model, and $ARI(X, Y)$ is the adjusted rand index between partitions X and Y, then I expect there to be some relationship of the nature:

$$\begin{aligned}
ARI(C^*, C') &\propto \log(N) \\
ARI(C^*, C') &\propto \sqrt{P_S} \\
ARI(C^*, C') &\propto -K \log(K) \\
ARI(C^*, C') &\propto \Delta_\mu \\
ARI(C^*, C') &\propto \frac{1}{\sigma} \\
ARI(C^*, C') &\propto \frac{\Delta_\mu K \log\left(\frac{N}{K}\right) \sqrt{P_S}}{\sigma}
\end{aligned}$$

I do not expect that the stated nature of these relationships is true, but the directionality of these relationships is expected to hold and something of the relative speed of improvement in resolving the true structure is intended to be indicated by the use of $\log(\cdot)$ and $\sqrt{\cdot}$. Note that the positive linear dependence on K is due to how I am defining Δ_μ. These statements also convey that for many of the variables it is relative values

that matter; for instance if Δ_μ increases we expect improved resolution of clusters, but if σ^2 also grows proportionally, then we would not expect the improvement to be at all as significant.

We will test:

1. The 2D Gaussian case (this is a sense-check);
2. The lack-of-structure case in 2 dimensions;
3. The large N, small P paradigm;
4. Increasing σ^2;
5. Increasing the number of irrelevant features;
6. The small N, large P case; and
7. Varying the proportion of the total population assigned to each sub-population.

Table 1: Longtable

N	P_s	P_n	K
100	2	0	5
100	0	2	1
10000	4	0	5
10000	4	0	50
10000	4	0	50
100	500	0	5
100	500	0	5
100	500	0	5
100	20	2	5
100	20	10	5
100	20	20	5
100	20	100	5
100	20	200	5
50	500	0	5
50	500	0	5
200	20	0	5

Law, Martin H, Anil K Jain, and Mário Figueiredo. 2003. "Feature Selection in Mixture-Based Clustering." In *Advances in Neural Information Processing Systems*, 641–48.