Consider two clusters A and B, where each has a normal distribution characterized by means, standard deviations, and a prior probability (P) of belonging to clusters A and B adds to 1, such that P(A), the probability of belonging to cluster A = 1 − P(B). The prior probability represents the expectation, and the calculation of the distribution parameters is the process of maximization.

**Processing Steps of the EM Algorithm**

The steps for processing the EM algorithm are as follows:

1. Start with initial guesses of the distributional parameters for each observation.
2. Use the initial guesses to calculate the cluster probabilities for each observation.
3. Use the calculated probabilities to re-estimate the parameters.
4. Go back to step 2 and do it again (until your time budget runs out).

The algorithm converges toward a fixed point but never gets there. But we can calculate the likelihood that the observation came from the data set, given the values for the parameters. The overall likelihood across all observations is the “goodness” of the clustering solution, and it increases during each iteration through the process. This likelihood may be only a “local” maximum (greater than all values near it), and there may be another maximum in another part of the probability landscape that is higher. The highest maximum across the entire probability landscape is the “global” maximum.

**V-fold Cross-Validation as Applied to Clustering**

The general idea of V-fold cross-validation as it is applied to clustering is to divide the overall sample into V folds, or randomly drawn (disjoint) subsamples. The same type of analysis is then successively applied to the observations belonging to the V − 1 folds (training sample), and the results of the analyses are applied to sample V (the sample or fold that was not used to estimate the parameters, build the tree, determine the clusters, etc.; i.e., this is the testing sample) to compute some index of predictive validity. The results for the V replications are aggregated (averaged) to yield a single measure of the stability of the respective model, i.e., the validity of the model for predicting new observations.

Cluster analysis is an unsupervised learning technique, and we cannot observe the (real) number of clusters in the data. However, it is reasonable to replace the usual notion (applicable to supervised learning) of “accuracy” with that of “distance”: In general, we can apply the V-fold cross-validation method to a range of numbers of clusters and observe the resulting average distance of the observations (in the cross-validation or testing samples) from their cluster centers (for k-means clustering); for EM clustering, an appropriate equivalent measure would be the average negative log-likelihood computed for the observations in the testing samples.

*Note:* The preceding discussion on k-means and EM clustering is based on Witten and Frank (2005).