A Model-based Semi-supervised Clustering Methodology

Jordan Yoder*
Department of Applied Mathematics and Statistics, Johns Hopkins University
and
Carey E. Priebe
Department of Applied Mathematics and Statistics, Johns Hopkins University

December 17, 2014

Abstract

We consider an extension of model-based clustering to the semi-supervised case, where some of the data are pre-labeled. We provide a derivation of the Bayesian Information Criterion (BIC) approximation to the Bayes factor in this setting. We then use the BIC to select the number of clusters and the variables useful for clustering. We demonstrate the efficacy of this adaptation of the model-based clustering paradigm through two simulation examples and a fly larvae behavioral dataset in which lines of neurons are clustered into behavioral groups.

Keywords: BIC, Machine learning, Behavioral data, Model selection, Mclust

*This work is partially funded by the National Security Science and Engineering Faculty Fellowship (NSSEFF), the Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.
1 Introduction

Clustering is the art of partitioning data into distinct and scientifically relevant classes by assigning labels to observations. Clustering is typically performed in an unsupervised setting, where none of the observed data initially have labels. In semi-supervised learning, some of the data have labels, but others do not; the goal is to classify the unlabeled data by assigning them to the same classes as the labeled data. In the gap between these two settings, there is semi-supervised clustering, in which some of the data have labels, and there may not be labeled data from every class. Furthermore, the total number of classes may be unknown. Just as in semi-supervised learning, leveraging the known labels allows for a more informed clustering.

There are many strategies for solving clustering problems. Some clustering procedures are based on heuristics that lack a principled justification, and many of the basic questions of clustering (e.g., how many classes there are) are often left to the intuition of the practitioner. Fraley and Raftery [2002] review model-based clustering, which recasts the task of clustering as a model selection problem, which is well-studied in the field of statistics. The main assumption in model-based clustering is that the data are drawn from one of $G$ distributions, where each distribution represents a cluster. Model selection can be accomplished by computing approximate Bayes factors for competing models with different numbers and/or types of components.

Raftery and Dean [2006] proposed a greedy algorithm for model-based clustering in the unsupervised setting; it used the BIC to choose the number of clusters and the features to consider while clustering. Murphy et al. [2010] adapted Raftery and Dean’s algorithm to semi-supervised learning. Semi-supervised clustering is between these two settings in terms of how much information is available, so we adapt Murphy et al.’s algorithm to incorporate all of the knowledge available. The main theoretical contribution of this paper is a calculation of approximate Bayes factors in the semi-supervised setting that facilitates this adaptation.

In this paper, we quickly review model based clustering in Section 2. We formally state our main theoretical result which generalizes the classical BIC to the semi-supervised case in Section 3. We give some illustrative simulation results in Section 4 and apply our methods
to the fly larvae behavioral dataset in Section 5. Finally, we conclude the paper with a discussion of the results and extensions to our method.

2 Review of Model Based Clustering

Assume that the data $X_1, X_2, \ldots, X_n$ are distributed i.i.d. according to a mixture model,

$$f_\theta(\cdot) = \sum_{k=1}^{G} \pi_k f_{\theta_k}(\cdot),$$

where $\theta_k$ are the parameters of the $k^{th}$ component of the mixture and $G$ is the total number of components. It can be shown that this is equivalent to the following generative process:

for each $i \in \{1, 2, \ldots, n\}$

1. draw $Z_i$ from multinomial $(\pi_1, \pi_2, \ldots, \pi_G)$
2. draw $X_i$ from $f_{\theta_{Z_i}}$.

If we consider each of the $G$ components as representing a single cluster, then the problem of clustering the data is that of unveiling each latent $Z_i$ from the generative process. The expectation-maximization (EM) algorithm can be used to find the maximum likelihood estimates for the parameters of the model and the posterior probabilities of cluster membership for each $X_i$ (cf. Dempster et al. [1977]).

With this formulation, we have a fully probabilistic clustering scheme. We can set the cluster of the $i^{th}$ observation to the maximum a posteriori estimate:

$$\hat{Z}_i = \arg\max_k \pi_k f_k(x_i).$$

Note that the posterior probability can give us a measure of confidence about the $i^{th}$ classification.

Here, we have used $f_{\theta_k}$ as a general density function, but it should be noted that the multivariate Gaussian distribution is the most common choice of distribution. We will eventually use an information criterion, such as the Bayesian Information Criterion (BIC), to assess the relative quality of different clusterings. Then, the number of clusters can be chosen using the BIC, which rewards model fit and penalizes model complexity.
By considering different constraints to the covariance matrices in the Gaussian components, we can reduce the number of parameters estimated, thus lowering the influence of the penalty term in the BIC. This allows for simpler clusters to be chosen over complex clusters. Celeux and Govaert [1995] considered a spectral decomposition of the $k^{th}$ component’s covariance matrix,
\[
\Sigma_k = \lambda_k D_k^T A_k D_k,
\]
where $\lambda_k$ represents the largest eigenvalue, $A_k$ is a diagonal matrix whose largest entry is 1, and $D_k$ is a matrix of the corresponding eigenvectors. The interpretation of each term as it relates to cluster $k$ is as follows: $\lambda_k$ represents the volume, $D_k$ the orientation, and $A_k$ the shape. By forcing one or more of these terms to be the same across all clusters, we can reduce the number of parameters to be estimated from $Gd + G\frac{d(d+1)}{2}$ in the unconstrained case ($\Sigma_k = \lambda_k D_k^T A_k D_k$) to $Gd + G - 1 + \frac{d(d+1)}{2}$ in the most constrained case ($\Sigma_k = \lambda D^T A D$).

We can also force additional constraints, such $D_k = I$ and/or $A_k = I$, leading to the simplest model: $\Sigma_k = \lambda I$ with only $G(d+1)$ parameters to estimate.

3 Methodology

The main theoretical contribution of this paper is to derive an adjustment to the BIC for the semi-supervised clustering setting. We will set forth a formalization of a more general setting, derive the BIC in this broader setting, and then apply our result to the special case of semi-supervised clustering. The modified BIC then represents a principled measure by which to choose the number of clusters and the variables for clustering when performing semi-supervised clustering.

Consider $n = \sum_{i=1}^{C} n_i$ independent random variables
\[
X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)} \overset{\text{i.i.d.}}{\sim} f_{\theta_1},
X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)} \overset{\text{i.i.d.}}{\sim} f_{\theta_2},
\vdots
X_1^{(C)}, X_2^{(C)}, \ldots, X_{n_C}^{(C)} \overset{\text{i.i.d.}}{\sim} f_{\theta_C},
\]
where $\theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1}$ and $\theta_j \in \Theta_j \subset \Theta_1$ for $j = 2, 3, \ldots C$ and
\[
(f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_C}) \in M = \{(f_{\theta^{(1)}}, f_{\theta^{(2)}}, \ldots, f_{\theta^{(C)}}) : \theta = (\theta^{(1)}, \ldots, \theta^{(C)}) \in \Theta_1 \times \Theta_2 \times \cdots \times \Theta_C\}.
\]
Collect all of the $X$’s into set $D$.

Model $M$, while more general than strictly necessary, encompasses the semi-supervised case. Consider the first group of $n_1$ random variables as those for which we do not have labels and each of the other $C - 1$ groups as those whose labels we know. Then, for a proposed total number of clusters $G \geq C - 1$, we assume

$$f_{\theta_1}(x) = \sum_{j=1}^{G} \pi_j \phi(x; \mu_j, \Sigma_j),$$

where $\phi(\cdot; \mu_j, \Sigma_j)$ is a multivariate normal pdf with mean $\mu_j$ and covariance matrix $\Sigma_j$, and $\sum_{j=1}^{G} \pi_j = 1$. Also, for each $k \in \{2, 3, \ldots, C\}$,

$$f_{\theta_k}(x) = \phi(x; \mu_{j_k}, \Sigma_{j_k}),$$

where the double subscript $j_k$ is to account for possible relabeling. It follows that

$$\theta_1 = ((\pi_1, \pi_2, \ldots, \pi_G), (\mu_1, \mu_2, \ldots, \mu_G), (\Sigma_1, \Sigma_2, \ldots, \Sigma_G)) \in \Theta_1,$$

where

$$\Theta_1 = \{(\pi_1, \pi_2, \ldots, \pi_G) : \sum_{i=1}^{G} \pi_i = 1, \pi_i \in [0, 1] \times \mathbb{R}^{d \times G} \times \{(\Sigma_1, \Sigma_2, \ldots, \Sigma_G) | \Sigma_i \succeq 0, \Sigma_i \in \mathbb{R}^{d \times d}\}.$$

Then, for each $k \in \{2, 3, \ldots, C\}$, $\Theta_k$ is the same as $\Theta_1$ except that the mixing coefficients $(\pi_1, \pi_2, \ldots, \pi_G)$ are constrained such that $\pi_{j_k} = 1$ and all other $\pi_i = 0$.

**Theorem 3.1.** The BIC approximation to the integrated likelihood of model $M$ is

$$2 \log(P(D|M)) \approx 2 \log(P(D|\hat{\theta}, M)) - d \log(n_1) = BIC_M,$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE).

We provide the proof in the appendix. Compare the above result to the classical BIC approximation:

$$BIC \approx 2 \log(P(D|M)) \approx 2 \log(P(D|\hat{\theta}, M)) - d \log(n).$$

The derivation of the BIC (adjusted and normal) involves an $O(1)$ term, which is dropped at the end. Due to this, the difference between the two BICs is a smaller penalty in the
semi-supervised case (i.e. $d \log(n_1)$ vs. $d \log(n)$). This can be interpreted as not penalizing more complicated cluster structures for the supervised data.

It is known that finite mixture models do not meet the regularity conditions for the BIC approximation used to approximate the integrated likelihood. However, Fraley and Raftery comment that it is still an appropriate and effective approximation in the paradigm of model-based clustering. Thus, we will not be too remiss in considering Theorem 3.1 as applicable when performing semi-supervised clustering. Because the model $M$ is appropriate for the semi-supervised case, we can use the modified BIC as a way to calculate Bayes factors for different number of total clusters $G$, subsets of variables to be included in the clustering, and parameterizations of $\Sigma_i$. The largest BIC value will therefore correspond to our chosen model for clustering. We will call the semi-supervised clustering algorithm using the BIC from Theorem 3.1 to make the aforementioned selections ssClust henceforth.

Actually calculating the BIC value from Theorem 3.1 involves computing the MLE. This is accomplished by the use of the EM algorithm. Both the E-step and the M-step are presented in the appendix.

4 Simulated Experiment

All relevant code for our algorithm is contained in the package ssClust (cf. Supplementary Materials). Further, each experiment has enough code provided to obtain one replicate.

4.1 Experiment 1

We construct an example that is difficult for many clustering algorithms in general. We draw 300 i.i.d. $d = 3$ dimensional data-points:

$$X_i \overset{i.i.d.}{\sim} \sum_{k=1}^{3} \frac{1}{3} \phi(\mu_k; \Sigma_k),$$

where $\mu_k = 1.35(k - 2)e$ for $k = 1, 2, 3$ and $e \in \mathbb{R}^3$ is a vector of all ones. For $k = 1, 3$, $\Sigma_k = J$, where $J \in \mathbb{R}^{3 \times 3}$ is the identity matrix. For $k = 2$, $\Sigma_2 = \frac{1}{2}J$. Thus, we have three spherical blobs of data which overlap (cf. Figure 1).
Figure 1: A pairs plot of one realization of the data when drawn according to the description in Section 4.1. Increasing amounts of the black squares will be pre-labeled for our semi-supervised clustering algorithm.

To demonstrate the efficacy of our algorithm, we vary the amount of pre-labeled class 2 data from 5 to 50 in increments of 5 and perform our semi-supervised clustering algorithm considering the following parameterizations for the covariance matrices: EEE, VVV, VVI, and VII (see Raftery and Dean [2006] for descriptions). Then, we use standard multivariate Gaussian model-based clustering considering the same parameterizations using Mclust for a control Fraley et al. [2012]. We will not perform any variable selection procedures in this experiment. We allow both algorithms to consider between two and five total clusters. To compare the performance of each algorithm, we compute the Adjusted Rand Index (ARI) for each method using only the $n_1$ unlabeled datapoints (cf. Hubert and Arabie [1985]).

We will briefly define the ARI. Consider the $n_1$ unlabeled data grouped into $G$ groups

$$X = \{X_1, X_2, \ldots, X_G\}$$

by an algorithm. Next consider the true grouping of the same data into $G^*$ groups:

$$Y = \{Y_1, Y_2, \ldots, Y_{G^*}\}.$$
Define

\[ ARI(X, Y) = \frac{\sum_{i,j} (|X_i \cap Y_j|)}{\frac{1}{2} \left( \sum_i \binom{|X_i|}{2} + \sum_j \binom{|Y_j|}{2} \right)} - \frac{\sum_i (|X_i|)}{\left( \sum_j \binom{|Y_j|}{2} \right)} - \frac{\sum_j (|Y_j|)}{\left( \sum_i \binom{|X_i|}{2} \right)} - \frac{1}{\binom{n}{2}} \sum_i (|X_i|) \sum_j (|Y_j|). \]

Finally, the above procedure was repeated 1000 times for each choice of number of pre-labeled data. The results are presented in Figure 2. We can see that our semi-supervised clustering method performs better than Mclust for all values of pre-labeled data even though we do not include the pre-labeled data in the ARI calculation. It should be noted that the ARI for Mclust increases as a function of the number of pre-labeled data because these data are the most commingled, so that excluding them from the ARI calculation improves Mclust performance.

![ARI vs. Number of Labels Known (MCR=1000)](image)

Figure 2: Average value of the ARI for 1000 Monte Carlo replicates for experiment 4.1. Error bars are ±2 standard errors.
4.2 Experiment 2

We now create a scenario to demonstrate the variable selection capabilities of our algorithm. We adapt the set-up in 4.1 to a more computationally tractable one, as Raftery and Dean’s algorithm requires many calls to Mclust. We draw 500 i.i.d. $d = 3$ dimensional data-points according to the distribution from 4.1. Next, we add an uninformative column of data to all data-points; we take 500 draws from a $\mathcal{N}(0, 1)$ distribution and treat these as the fourth feature. The hope is that we will be able to identify this feature and discard it. To this aim, we use an adaptation of clustvarsel, the CRAN implementation of Raftery and Dean’s algorithm, which uses our semi-supervised clustering algorithm in lieu of Mclust to choose the variables to use for clustering. We allow both algorithms to consider between two and four total clusters and the following parameterizations of covariance matrices: VVV and VII. Then, we calculate the ARI of the clustering associated with the chosen variables. Similarly, we use clustvarsel to select the variables for Mclust and report the ARI.

The above procedure was repeated 500 times for each choice of number of pre-labeled data. The results are presented in Figure 3. We can see that our semi-supervised clustering method performs generally better than Mclust, but that both algorithms performed worse than in 4.1. Note that it is reasonable for the ARI values to be lower than in the previous experiment, because of the extra difficulty of the additional useless dimension.
5 Real Data Example

5.1 Using Semi-Supervised Data to Recover Labels

In one of the motivating applications for this work, classes of neurons in *Drosophila* larvae are controlled using optogenetics, cf. opt [2011]. In Vogelstein et al. [2014], they observe the reactions of the affected larvae to stimuli in high-throughput behavioral assays. The goal is to determine which classes of neurons cause similar changes in behavior when deactivated.

We initially collected data on $n = 37780$ larvae grouped into $b = 2062$ dishes. By changing the optogenetic procedure, $\ell = 11$ known lines are created, pbd1, 38a1, 61d0, ppk1, 11f0, pbd2, 38a2, pbd3, ppk2, iav1, and 20c0. Of these lines, we discarded the larvae in pbd3 and ppk2 because they had less than 40 larvae each. Further, we discarded all larvae with an unknown line. After curating the data, we now have $n = 7730$ larvae. Each larva is observed while responding to various stimuli. Vogelstein et al. [2014] expand on the methods of Priebe [2001] and Priebe et al. [2004] and describe how the observations are embedded into $\mathbb{R}^d$, where here $d = 30$. We then used the method presented in Zhu and
Ghodsi 2006 to select the elbow of the scree plot to further reduce the data to $d = 14$ dimensions. We did not perform any additional feature selection. Thus, this experiment is most similar to Experiment 4.1.

For each Monte Carlo replicate, we use a small subset of the data where the line was known (101 randomly chosen animals from each of the 9 remaining lines) along with $m = 0, 1, 2, \ldots 10$ pre-labeled data randomly chosen from the 101 animals in each line. Then, we compute the ARI against the line type and compared this to the Mclust results. We find that a single animal per line significantly improves the clustering results, as expected (cf. Figure 4). Further, we can see that there are diminishing returns on additional supervision starting at 3 supervised examples per line.

Figure 4: Average value of the ARI for 500 Monte Carlo replicates for experiment 5.1. Error bars are ±2 standard errors.

This experiment is most comparable to 2 in the sense that we do not perform variable selection concurrently with clustering, as in 3.

5.2 Differentiating Lines

Vogelstein et al. 2014 posed and answered questions of the form “Is line X different
than line Y (in terms of behaviotypes)?” As an illustrative example, we will perform a similar analysis comparing the ppk1 and pbd1 lines but will additionally incorporate some supervision. Here, our interpretation of the clusters will shift from lines to behaviotypes. Our proposed procedure for distinguishing between lines is as follows:

1. Sample 101 animals from each line
2. Label 3 animals from each line according to a labeling strategy (see below)
3. Cluster the animals using ssClust and Mclust into between 2 and 12 clusters using the parameterizations EEE, VVV, VII, and EII.
4. Collect the results and construct empirical probability of cluster membership for each line
5. Compute the Hellinger distance between the two lines to be compared and store this as statistic $H$
6. Simulate the distribution of $H$ under the null hypothesis that the lines are the same by permuting the labels and computing the Hellinger distance $H_i$ for $i = 1, 2, \ldots, B$ for some large integer $B$.
7. Return an empirical p-value based on steps (5) and (6).

In item (2) we did not specify a labeling strategy in detail. We propose a strategy that is reasonably realistic to execute for our particular dataset. Vogelstein et al. [2014] used a hierarchical clustering paradigm in which the first few layers were visually identifiable by watching the worms. Thus, by using their labels from an early layer (layer 2, with 4 clusters total), we have a plausible level of supervision for a human to have performed. Specifically, we sample at random a label from the true labels among a line with weights proportional to the counts of each label in that line. Using that label, we sample 3 worms of that label in that line to be the supervised examples. Next, for the other line, we sample a different label, and 3 examples with that true label.

We see that based on all three labeling strategies that both ssClust and Mclust are able to corroborate the results from Vogelstein et al. [2014] even with small amounts of data: ppk1 and pbd1 are statistically different (p-value $\approx 1e^{-4}$ for both for 500 MC replicates).
We now show that ssClust can answer these questions “sooner” than Mclust. That is, the p-value will be below the significance level with fewer unsupervised examples. To quantify this concept, we introduce the “answering time” for algorithm A:

$$\tau_A := \min_{n \in \{2, 3, 4, \ldots, 30\}} n$$

such that

$$\prod_{q=n}^{30} \mathbb{1}\{pVal(D_k) \leq \alpha\} = 1,$$

where here we use the notation

$$D_k := \{X_{1,k}, \ldots, X_{q,k}, (X_{99,k}, Y_{99,k}), \ldots, (X_{101,k}, Y_{101,k}) : k \in \{1, 2\}\}$$

to be the labeled and unlabeled data.

The p-value constraint bears some explanation; it says that we require a significant p-value for all datasets at least as large as with n unsupervised examples per line. Note that here we assume our datasets are nested and that they all use the same supervised examples. Since the answering time will be dependent on the datasets used, we perform a Monte Carlo simulation with 500 random sequences of datasets and report the answering times for both ssClust and Mclust (cf. Figure 5). The median answering time for ssClust is evidently lower than for Mclust (pval = 1.6e-4 for paired Wilcoxon signed-rank test).
6 Conclusion

Previously explored approaches to semi-supervised clustering include a modified K-means, which can be seen as our algorithm constrained to spherical and identical covariance matrices without the adjusted BIC for model selection (cf. Basu et al. [2002]). Others have used latent variables representing clusters and cluster-dependent distributions to give a probabilistic approach to a slightly different problem, where instead of labels being known, only pairwise constraints in the form of must-link and cannot-link are given (cf. Basu et al. [2004]). Wagstaff et al. [2001] applied an appropriately modified K-means to this problem. Finally, Cohn et al. [2003] and Xing et al. [2002] offer a different approach to semi-supervised clustering involves training a measure to account for constraints or labels.

The main contribution of this paper over previous works is presenting a probabilistic framework for semi-supervised clustering and deriving an information criterion consistent with the framework to allow selection of number of clusters and/or clustering variables. With the corrected BIC, we found that in the simulated examples and fly larvae dataset our method outperformed the most comparable method, Mclust. In the fly dataset, ssClust
was able to recover lines better and yielded a lower answering time more often for the question of whether two particular lines were different, behaviorally. This indicates that incorporating even a small amount of information can help guide the clustering process.

Future areas of research will involve handling the more nuanced must-link and cannot-link constraints, which are flexible enough to encompass the labels-known problem we have explored in this paper.

7 Appendix

Assume the data are distributed according to a member of model $M$ described in Section 3.

Proof of Theorem 3.1

Consider the integrated likelihood:

$$P(D) = \int P(D|\theta, M)P(\theta|M)d\theta,$$

where $P(\cdot)$ is a probability, pmf, or pdf where appropriate. Let $Q(\theta) = \log (P(D|\theta, M)P(\theta|M))$, the log of the posterior likelihood. Suppose that the posterior mode exists, say $\bar{\theta}$.

A second order Taylor expansion about $\bar{\theta}$ gives

$$Q(\theta) = Q(\bar{\theta}) + (\theta - \bar{\theta})^T \nabla Q(\bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})^T \nabla^2 Q(\bar{\theta})(\theta - \bar{\theta}) + o(\|(\theta - \bar{\theta})\|^2).$$

By the first order optimality necessary conditions, we know $\nabla Q(\bar{\theta}) = 0$. Note $o(\|(\theta - \bar{\theta})\|^2) \rightarrow 0$ faster than $\|(\theta - \bar{\theta})\|^2$ as $\theta \rightarrow \bar{\theta}$. By ignoring the last term, we will approximate $Q(\theta)$ with the truncated Taylor expansion:

$$Q(\theta) \approx Q(\bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})^T \nabla^2 Q(\bar{\theta})(\theta - \bar{\theta}).$$
Recalling (1), we have
\[ P(D|M) = \int \exp \left( \log(P(D|\theta, M)P(\theta|M)) \right) d\theta \]
\[ = \int \exp(Q(\theta)) d\theta \]
\[ \approx \int \exp \left( Q(\bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})^T \nabla^2 Q(\bar{\theta})(\theta - \bar{\theta}) \right) d\theta \]
\[ = \exp(Q(\bar{\theta})) \int \exp \left( \frac{1}{2}(\theta - \bar{\theta})^T \nabla^2 Q(\bar{\theta})(\theta - \bar{\theta}) \right) d\theta \]
\[ = \exp(Q(\bar{\theta})) \int \exp \left( -\frac{1}{2}(\theta - \bar{\theta})^T (-\nabla^2 Q(\bar{\theta}))(\theta - \bar{\theta}) \right) d\theta. \]

Recognize the integral as proportional to the density of a multivariate Guassian with mean \( \bar{\theta} \) and covariance \(-\nabla^2 Q(\bar{\theta})\). Let \( H = -\nabla^2 Q(\bar{\theta}) \). Then, we have
\[ P(D|M) \approx \exp \left( Q(\bar{\theta}) \right) \left( 2\pi \right)^{d/2} \det(H^{-1})^{1/2}, \]
where \( d \) is number of free parameters in \( \theta \). If the conditions in Proposition 3.4.3 in [Bickel and Doksum 2001] hold, we have for \( n \) sufficiently large, \( H \approx I(\bar{\theta}) \), where \( I(\bar{\theta}) \) is the Fisher information matrix.

Taking \( 2 \log(\cdot) \) of both sides, (2) becomes
\[ 2 \log(P(D|M)) \approx 2Q(\bar{\theta}) + d\log(2\pi) + \log(\det(I(\bar{\theta})^{-1})) \]
\[ = 2\log(P(D|\bar{\theta}, M)) + 2\log(P(\bar{\theta}|M)) + d\log(2\pi) - \log(\det(I(\bar{\theta}))). \]

Now, we must calculate \(-\log(\det(I(\bar{\theta}))\). Define \( n \equiv \sum_{i=1}^{C} n_i \). Observe
\[ I(\theta) = \text{Var} \left[ \frac{\partial}{\partial \theta} \log(p(X, \theta)) \right] \]
\[ = \text{Var} \left[ \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(p(X_i, \theta)) \right] \]
\[ = \sum_{i=1}^{C} n_i I(\theta_i) \text{ by independence.} \]

Henceforth, we will use \( I_j \) to denote \( I(\theta_j) \). We will assume that \( I_1 \) is positive definite, so that it can be written as
\[ I_1 = S_1^T S_1 \]
for some non-singular matrix $S_1$. Let $J$ denote the $d$-dimensional identity matrix. For notational purposes, let
\[
n^* := \max_{2 \leq j \leq C} (n_j).
\]
Observe
\[
\det \left( \sum_{i=1}^{C} n_i I(\theta_i) \right) = \det \left( S_1^T (n_1 J + (S_1^T)^{-1} \left( \sum_{i=2}^{C} n_i I(\theta_i) \right) S_1^{-1}) S_1 \right)
\]
\[
= \det(I_1) \det \left( n_1 J + n^* (S_1^T)^{-1} \left( \sum_{i=2}^{C} \frac{n_i}{n^*} I(\theta_i) \right) S_1^{-1} \right)
\]
\[
= \det(I_1) n_1^d \det \left( J + \frac{n^*}{n_1} (S_1^T)^{-1} \left( \sum_{i=2}^{C} \frac{n_i}{n^*} I(\theta_i) \right) S_1^{-1} \right)
\]
\[
= \det(I_1) n_1^d \det \left( J + \frac{n^*}{n_1} B \right),
\]
where $B = (S_1^T)^{-1} (\sum_{i=2}^{C} \frac{n_i}{n^*} I(\theta_i)) S_1^{-1}$. It should be noted that $B$ is a positive semi-definite matrix, as it is a $*$-transform of a sum of positive semi-definite matrices, so that Sylvester’s Theorem states that $B$ has the same inertia as a positive semi-definite matrix. Suppose that $\frac{n^*}{n_1} \leq M$ for some $M$ independent of $n$.

Next, we would like to find a bound for $\det \left( J + \frac{n^*}{n_1} B \right)$ that is independent of $n$. For any eigenvalue of $J + \frac{n^*}{n_1} B$, say $\lambda$, we have by Weyl’s theorem
\[
1 \leq \lambda \leq 1 + \frac{n^*}{n_1} \|B\|_2 \leq 1 + M \|B\|_2.
\]
If we can bound $\|B\|_2$ independently of $n$, we have our bound on the determinant. To this aim, observe
\[
\|B\|_2 = \max_{\{x \in \mathbb{R}^d : \|x\|_2 = 1\}} x^T B x
\]
\[
\leq \max_{\{x \in \mathbb{R}^d : \|x\|_2 = 1\}} \|S_1^{-1}\|_2^2 \left( \sum_{j=2}^{C} \frac{n_j}{n^*} \|I_j\|_2^2 \right)
\]
\[
\leq \max_{\{x \in \mathbb{R}^d : \|x\|_2 = 1\}} \|S_1^{-1}\|_2^2 \left( \sum_{j=2}^{C} \|I_j\|_2^2 \right),
\]
which is independent of $n$. Let $M_2 \equiv \max_{x \in \mathbb{R}^d, \|x\|_2 = 1} \|S_1^{-1}\|_2^2 \left( \sum_{j=2}^C \|I_j\|_2^2 \right)$. Then, we have

$$1 \leq \det(J + \frac{n^*}{n_1} B) = \prod_{m=1}^d \lambda_m \left( J + \frac{n^*}{n_1} B \right) \leq (1 + M(M_2))^d.$$ 

Hence,

$$\log(\det(I(\theta))) = \log(\det(I_1)) + d \log(n_1) + \log \left( \det(J + \frac{n^*}{n_1} B) \right) \leq \log(\det(I_1)) + d \log(n_1) + O(1),$$

where the $O(1)$ term indicates that $\log \left( \det(J + \frac{n^*}{n_1} B) \right)$ is bounded as $n \to \infty$.

When the posterior mode is nearly or is equal to the MLE, as is the case when the prior on $\theta$ is uniform and $\Theta$ is finite (cf. Bickel and Doksum pp. 114), substitute $\hat{\theta}$, the MLE, for $\bar{\theta}$, the posterior mode. Then, we have

$$2 \log(P(D|M)) \approx 2 \log(P(D|\hat{\theta}, M)) + 2 \log(P(\hat{\theta}|M)) + d \log(2\pi) - \log(\det(I_1(\hat{\theta})))$$

$$= 2 \log(P(D|\hat{\theta}, M)) - d \log(n_1) + O(1) \text{ using equation (2).}$$

Note that the $O(1)$ term’s importance diminishes in the limit as $n \to \infty$, because the other terms are unbounded. If we drop it, we have our derivation of the adjusted BIC.

**E-Step for semi-supervised clustering.**

Let $Z_{i,k} = 1\{X_i \text{ is in cluster } k\}$ denote the (hidden) cluster memberships, where $1\{\cdot\}$ is the indicator function. The log-likelihood of the parameters given the data is

$$l(\theta, Z_{i,k}|D) = \sum_{i=1}^n \sum_{k=1}^G Z_{i,k} \log(\pi_k \phi(X_i; \mu_k, \Sigma_k)).$$

Note that for the $n - n_1$ data not in the first group, we know the true $Z_{i,k}$, but for the $n_1$ unlabeled data we do not. Thus, the E-step of the EM algorithm is

$$\mathbb{E}[l(\theta, Z_{i,k}|D)] = \sum_{i=1}^{n_1} \sum_{k=1}^G \mathbb{E}[Z_{i,k}|D] \log(\pi_k \phi(X_i; \mu_k, \Sigma_k)) + \sum_{k=2}^C \sum_{i=n_{k-1}+1}^{n_k} \log(\phi(X_i; \mu_j, \Sigma_j)).$$

By the definition of $Z_{i,k}$ and Bayes Theorem, we have for all $i = 1, 2, \ldots, n_1$ that

$$\mathbb{E}[Z_{i,k}|D] = P(Z_{i,k} = 1|D) = \frac{\pi_k \phi(X_i; \mu_k, \Sigma_k)}{\sum_{j=1}^G \pi_j \phi(X_i; \mu_j, \Sigma_j)}.$$
M-Step for semi-supervised clustering.
Define $\rho_{i,k} = \mathbb{E}[Z_{i,k} | D]$. We must maximize $\mathbb{E}[(\theta, Z_i, k | D)]$ with respect to the mixing coefficients $\pi_k$, the mean vectors $\mu_k$, and the covariance matrices $\Sigma_k$ for $k = 1, 2, \ldots, G$. It can be shown that the update equations for $\pi_k$ are

$$\pi_k = \frac{\sum_{i=1}^{n_1} \rho_{i,k}}{n_1}.$$

Recalling that $\rho_{i,k} = 1 \{ X_i \text{ is in cluster } k \}$ for $i = n_1 + 1, n_1 + 2, \ldots n$, we can now use the standard update equations for $\mu_k$ and $\Sigma_k$. Specifically,

$$\mu_k = \frac{\sum_{i=1}^{n} \rho_{i,k} X_i}{\sum_{i=1}^{n} \rho_{i,k}},$$

the weighted average of the data with weights equal to the posterior probabilities of cluster membership. The exact equations for the update equations for $\Sigma_k$ varies based on the parsimonious parameterization chosen (e.g. EEE). Celeux and Góvert derive the closed and iterative forms for a variety of parameterizations in Celeux and Govaert [1995].

References

Method of the year 2010. Nat Meth, 8(1):1–1, 01 2011. URL [http://dx.doi.org/10.1038/nmeth.f.321](http://dx.doi.org/10.1038/nmeth.f.321).

Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In Proceedings of 19th International Conference on Machine Learning (ICML-2002). Citeseer, 2002.

Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 59–68. ACM, 2004.

Peter J. Bickel and Kjell A. Doksum. Mathematical Statistics, volume I. Prentice-Hall, Inc., 2001.

Gilles Celeux and Gérard Govaert. Gaussian parsimonious clustering models. Pattern recognition, 28(5):781–793, 1995.
David Cohn, Rich Caruana, and Andrew McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 4(1):17–32, 2003.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, University of Washington, Department of Statistics, June 2012.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Thomas Brendan Murphy, Nema Dean, and Adrian E Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The annals of applied statistics*, 4(1):396, 2010.

Carey E. Priebe. Olfactory classification via interpoint distance analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(4):404–413, 2001.

Carey E Priebe, David J Marchette, and Dennis M Healy Jr. Integrated sensing and processing decision trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):699–708, 2004.

Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
Joshua T Vogelstein, Youngser Park, Tomoko Ohyama, Rex A Kerr, James W Truman, Carey E Priebe, and Marta Zlatic. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science*, 344(6182):386–392, 2014.

Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584, 2001.

Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512, 2002.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics &amp; Data Analysis*, 51(2):918–930, 2006.