Bayesian Mixture Models For Semi-Supervised Clustering

Amine Echraibi, Joachim Flocon-Cholet, Stéphane Gosselin, Sandrine Vaton

Bayesian Mixture Models For Semi-Supervised Clustering

Amine Echraibi[1], **Joachim Flocon-Cholet**[1], **Stéphane Gosselin**[1], and **Sandrine Vaton**[2]

[1]Orange Labs, France. {amine.echraibi, joachim.floconcholet, stephane.gosselin}@orange.com
[2]IMT Atlantique, France. sandrine.vaton@imt-atlantique.fr

Abstract. In most real-world applications of clustering, data is partially labeled by an expert. Classical clustering approaches have been extensively studied in the presence of partial labels, however little work has been done to treat the general case of Bayesian mixture models. In this paper, we propose a new approach to perform semi-supervised clustering using parametric and non parametric mixture models. We show how our approach generalizes mixture models with different types of emission distributions and priors under the same theoretical framework for semi-supervised clustering. The partial labels intervene in the clustering in the form of a Hidden Markov Random Field (HMRF) that introduces a penalty if the partial labels are not respected. We demonstrate how to perform inference in both the finite and infinite case with priors on the mixture components and the parameters using variational inference. Our experimental evaluations on synthetic data show how the method can leverage the partial labels to choose the correct clustering and the correct number of clusters. We also show that by introducing a small fraction of partial labels our method improves the clustering accuracy and outperforms a strong baseline in the literature on benchmark datasets.

1 Introduction

Semi-supervised learning gained considerable interest from researchers recently due to the availability of large amounts of unlabeled data and a small fraction of labeled data [7]. A sub task of semi-supervised learning is semi-supervised clustering i.e clustering on partially labeled datasets. These partial labels are usually set by an expert who spent considerable time working on the data and developing some knowledge about the domain [1]. Although this knowledge represented by the partial labels is limited, it could be useful during the clustering analysis. These partial labels can help to identify the number of clusters or the correct view of the clustering, for example.

In the semi-supervised learning literature, two approaches to assign classes to samples with prior knowledge have been extensively studied. The first approach supposes that all the different classes are present in the labeled part of the dataset, thus the number of classes is known. The model then attempts to propagate the labels to the unlabeled set, by respecting some notion of similarity. Two noteworthy methods of such kind are Transductive Support Vector Machines [10] and label propagation [4] and [17]. The Transductive Support Vector Machine model leverages the unlabeled data in the learning of the decision boundary between classes unlike classical support vector machines where only the labeled data is used. Label propagation, on the other hand, is a graph based approach where the labeled and unlabeled data points represent nodes of the graph, and edges represent similarities between the nodes. The known labels are then propagated through the graph to label the unlabeled nodes. However, in some real world applications of data mining and information retrieval, the number of possible clusters is unknown and not all clusters are partially labeled. Therefore, we can not use these methods.

The second approach is based on the classical KMeans clustering algorithm, where constraints are added in order to guide the clustering, and improve the performance [13]. The constraints are constructed from the partial labels, must-link constraints assure that two samples must have the same label, and cannot-link constraints assure that two samples shouldn't be labeled the same. [2] proposed a Hidden Markov Random Field (HMRF) formulation of this model and later introduced distance learning to identify relevant features to the clustering [3]. In this approach it is possible to find a hidden cluster not present in the partial labels. However, because it is based on the KMeans algorithm, the number of clusters must be known or some sort of model selection needs to be applied to estimate it. For a detail review of these approaches, we refer the reader to [7].

In this paper, we generalize the semi-supervised clustering approach with Hidden Markov Random Fields to the broader range of Bayesian mixture models. This is particularly interesting from a practical perspective, since in some cases we may need to consider various types of probability distributions for clustering, such as categorical distributions or Dirichlet priors for text clustering. Furthermore, by considering Bayesian priors on the mixture weights such as Dirichlet distributions [11] or Dirichlet process priors [6], we can estimate the number of clusters from the data itself.
Our contributions[1] in this paper are the following:

- We propose a general framework to perform semi-supervised clustering with Bayesian mixture models, by introducing a Hidden Markov Random Field prior on the hidden class variables.
- We formulate the pairwise potentials of the HMRF using distances between distributions, which allows for the same definition to be applied to different kinds of mixture models.
- We show how to perform inference on the model using variational inference and the mean field approximation [14].
- We show, in the experiments and results, that our model outperforms a strong baseline on benchmark datasets. We also show that it is capable of identifying the correct view of the clustering and the correct number of clusters, and how it can also improve performance of classical unsupervised Bayesian mixture models.

[1] A public implementation is available at: `https://git.io/Je6kb`

The remainder of this paper is organized as follows. First, we present the theoretical framework of semi-supervised clustering in the general case of exponential family mixture models. We develop the Hidden Markov Random Field formulation that allows for a generalization over different types of probability distributions. Then, we show how to use variational approximation methods to perform inference on the model. Finally, we present some didactic examples to give an intuition on how the model behaves in simple situations, and we show how the semi-supervision can improve the clustering performance of classical Bayesian mixture models.

2 Semi-Supervised Bayesian Mixture Models

2.1 Notations

First we introduce some notations that we will use throughout the paper. Let us consider a dataset $\mathbf{x}_{1:N} = \{\mathbf{x}_n\}_{n=1}^N$ composed of N i.i.d. samples of the random variable X. Let $\mathbf{z}_{1:N}$ be N latent random variables where $\mathbf{z}_n$ represents the class of sample $\mathbf{x}_n$. We denote by $\theta_{1:K}$ the random variables representing the parameters of the emission distributions of the mixture model, and by π the random variables representing the mixture weights, where $K \in \mathbb{N} \cup \{\infty\}$. We denote by p_{θ_k} the prior on the kth parameter θ_k. A Bayesian mixture model is defined for all k and n as:

$$\begin{aligned} \pi &\sim \text{Dir}(K, \alpha) \text{ or GEM}(\eta) \\ \theta_k &\sim p_{\theta_k}(\cdot) \quad \text{prior on } \theta_k \\ \mathbf{z}_n|\pi &\sim \text{Cat}(\cdot|\pi) \\ \mathbf{x}_n|\mathbf{z}_n = k, \theta &\sim p_X(\cdot|\mathbf{z}_n = k, \theta) \quad \text{emission distribution} \end{aligned}$$

where the mixture weights π follow a Dirichlet distribution in the case of a finite mixture model, or a Dirichlet process in the case of an infinite mixture. In this case, we adopt the *stick-breaking construction* definition [12] where $\pi \sim \text{GEM}(\eta)$ is equivalent to:

$$\begin{aligned} \forall k \quad \beta_k &\sim \text{Beta}(1, \eta) \\ \pi_k &= \beta_k \prod_{l=1}^{k-1}(1 - \beta_l) \end{aligned}$$

In the semi-supervised case, some instances of the dataset are labeled. Let us denote by $l_{1:N}$ the partial labels, where $l_n \in \{1, ..., K\}$ if $\mathbf{x}_n$ is labeled, $l_n = -1$ otherwise. For the rest of the paper, to simplify notation, we simply write for the probability distribution of a random variable X: $p_X(x) = p(x)$.

2.2 Supervision in the form of a HMRF

In order to incorporate the information provided by the partial labels into the model, we define a pairwise Hidden Markov Random Field over $\{\mathbf{z}_{1:N}, \mathbf{x}_{1:N}\}$ (Figure 1). The HMRF introduces a statistical dependency between the random variables $\mathbf{z}_n$ with the same label l_n:

$$\mathbf{z}_n \text{ neighbor of } \mathbf{z}_m \iff l_n = l_m$$

and each $\mathbf{x}_n$ is independent of $\mathbf{x}_m$ given $\mathbf{z}_n$ and θ. We write $n \sim m$ if $\mathbf{z}_n$ is a neighbor of $\mathbf{z}_m$, The neighborhood of the n^{th} sample is defined as :

$$\mathcal{N}_n = \{m \in \{1, ..., N\} \backslash \{n\} \text{ s.t } m \sim n\}$$

The statistical dependency introduced by the neighborhood is later used to construct a joint prior over $\mathbf{z}_{1:N}$ in such a way that samples in the same neighborhood should be assigned the same label. The definition of the neighborhood can be adapted to other formulations of semi-supervised clustering, such as, the must-link constraint and cannot-link constraints used in [3].

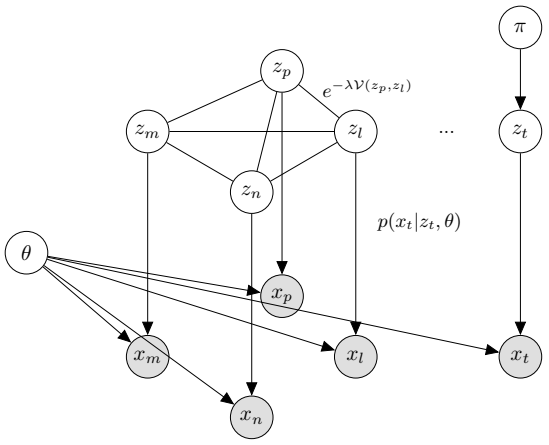

Figure 1: Overview of the probabilistic graphical model. The unlabeled samples are treated as a classical mixture model where x_t depends on θ and z_t, and z_t depends on π. The partially labeled samples create dependencies between the hidden variables $\mathbf{z}_{1:N}$. For each connected pair (z_p, z_l) we associate a factor $e^{-\lambda\mathcal{V}(z_p,z_l)}$.

2.3 Semi-Supervised Mixture Models

The Hidden Markov Random Field introduces dependencies between the latent variables $\mathbf{z}_{1:N}$. The generative process of the semi-supervised mixture model in this case becomes:

$$\begin{aligned} \pi &\sim \text{Dir}(K, \alpha) \text{ or GEM}(\eta) \\ \theta_k &\sim p_{\theta_k}(\cdot) \\ \mathbf{z}_{1:N}|\pi &\sim p_{\mathbf{z}_{1:N}}(\cdot|\pi) \\ \mathbf{x}_n|\mathbf{z}_n = k, \theta &\sim p_X(\cdot|\mathbf{z}_n = k, \theta) \end{aligned}$$

where $p(\mathbf{z}_{1:N}|\pi)$ is the joint distribution over the latent variables defined by the HMRF. According to the Hammersley–Clifford theorem [8], this joint distribution can be written as :

$$p(\mathbf{z}_{1:N}|\pi) = \frac{1}{\Gamma} \prod_{n=1}^{N} \pi_{\mathbf{z}_n}^{\mathbb{1}[\mathcal{N}_n = \emptyset]} \prod_{n \sim m} e^{-\lambda\mathcal{V}(\mathbf{z}_n, \mathbf{z}_m)}$$

where $\mathcal{V}$ represents the pairwise potentials of the HMRF, and λ is a scaling parameter that can be tuned empirically. Given how the HMRF is factored, we can write the normalization constant Γ as (proof in appendix B):

$$\Gamma = \sum_{\mathbf{z}_n, \forall n \text{ s.t } \mathcal{N}_n \neq \emptyset} \prod_{n \sim m} e^{-\lambda\mathcal{V}(\mathbf{z}_n, \mathbf{z}_m)}$$

The main idea behind the HMRF prior, is that neighboring samples should have the same label, otherwise the log-likelihood of the data is penalized by a factor proportional to the pairwise potentials. Other works in the literature proposed various definitions of the potentials, these definitions are often domain or data specific [16]. Others attempted to learn these potentials by introducing a parameterized distance [3]. In the following section, we show how to perform inference on the model, and we propose a definition for the potentials $\mathcal{V}$ constructed from the variational approximating distributions. Therefore this definition can be applied to all types of mixture models.

2.4 Variational Approximation

In order to perform inference on the model, we need to compute or estimate:

$$p(\mathbf{z}_{1:N}, \theta, \pi | \mathbf{x}_{1:N}) \propto p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi) \propto \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \theta) p(\mathbf{z}_{1:N}|\pi) p(\theta) p(\pi)$$

Performing exact inference on this model is intractable. In order to approximate the joint distribution $p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi)$, we use variational inference and the mean field approximation:

$$q(\mathbf{z}_{1:N}, \theta, \pi) = \prod_{n=1}^{N} q(\mathbf{z}_n) \prod_{k=1}^{T} q(\theta_k) q(\pi)$$

where T is a truncation level for the number of clusters [6]. The optimal approximate distribution q^* satisfies:

$$q^* = \arg\min_q \mathbb{D}_{KL}\left[q||p\right] \tag{1}$$

Solving equation (2) leads to the following mean field update equations:

$$\begin{aligned} \log q(\mathbf{z}_n) &= \text{const} + \mathbb{E}_{\{\mathbf{z}_{-n}, \theta, \pi\} \sim q}\left[\log p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi)\right] \\ \log q(\theta_k) &= \text{const} + \mathbb{E}_{\{\mathbf{z}_{1:N}, \theta_{-k}, \pi\} \sim q}\left[\log p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi)\right] \\ \log q(\pi) &= \text{const} + \mathbb{E}_{\{\mathbf{z}_{1:N}, \theta\} \sim q}\left[\log p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi)\right] \end{aligned}$$

By substituting the expression of $p(\mathbf{z}_{1:N}, \mathbf{x}_{1:N}, \theta, \pi)$ in the previous equations we have:

$$q(\mathbf{z}_n) = \text{Cat}(\mathbf{z}_n; \phi_n)$$

where:

$$\log \phi_{nk} = \text{const} + \mathbb{E}_{\theta \sim q}[\log p(\mathbf{x}_n|\mathbf{z}_n = k, \theta)] + \mathbb{1}[\mathcal{N}_n = \emptyset] \mathbb{E}_{\pi \sim q}[\log \pi_k] - \lambda \sum_{m \in \mathcal{N}_n} \sum_{l=1}^{T} \phi_{ml} \mathcal{V}(k, l) \tag{2}$$

and for the mixture parameters:

$$\log q(\theta_k) = \text{const} + \sum_{n=1}^{N} \phi_{nk} \log p(\mathbf{x}_n|\mathbf{z}_n = k, \theta) + \log p(\theta_k)$$

$$\log q(\pi) = \text{const} + \sum_{n=1}^{N} \sum_{k=1}^{T} \mathbb{1}\left[\mathcal{N}_n = \emptyset\right] \phi_{nk} \log \pi_k + \log p(\pi)$$

Usually the prior and the emission distribution are conjugate so $q(\theta_k)$ is in the same exponential family as $p(\theta_k)$. Therefore we can derive fixed point update equations for the variational parameters of $q(\pi)$ and $q(\theta_k)$ in the finite and infinite case. However in order to have a closed form for the update equation of the local variational parameter ϕ_{nk}, we need to propose a definition of $\mathcal{V}(k, l)$, which can be computed in close form, and have the property of introducing a penalty in the form of a distance between the clusters. Hence:

$$\mathcal{V}(k, l) \triangleq \mathbb{D}_{KL}^{sym}\left[q(\theta_k)||q(\theta_l)\right] \tag{3}$$

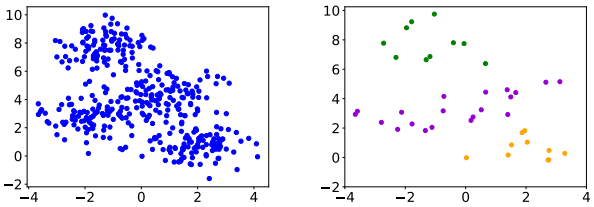

Figure 2: Samples from 4 Gaussians (left). Partial labels (right).

2.5 Intuition behind the model

The mean field update equations show that the model behaves the same as a classical unsupervised Bayesian mixture model, except in the case of the mean field update for ϕ_{nk} . The first two terms of equation (3) are the same as a classical unsupervised model, where $q(z_n = k)$ depends on the emission probability and the mixture weights. The third term however, introduces the supervision captured in the Hidden Markov Random Field, where for each sample m in the neighborhood of n a penalty is introduced proportional to $\phi_{ml}\mathcal{V}(k, l)$. Therefore if a sample m in the neighborhood takes a different label: $l \neq k$ ($\phi_{ml} \approx 1$) a penalty $\mathcal{V}(k, l)$ equivalent to the distance between the approximating distributions $q(\theta_k)$ and $q(\theta_l)$ forces the samples in the same neighborhood to have the same labels. Which in turn constrains the mixture parameters in the following fixed point updates to respect the partial labels.

3 Experiments and Results

We verify through experiments on synthetic data that the model is capable of identifying the correct number of clusters and the correct view of the clustering by leveraging information from the partial labels. We generate 400 samples from a mixture of Gaussian distributions of two dimensions. We label a random subset of the samples and we set them as the partial labels. In all the following experiments we adopt the Semi-Supervised Dirichlet Process Gaussian Mixture Model presented in appendix A. In our implementation we initialize the values of the local variational parameters ϕ_{nk} using the distance from the centers of a fitted KMeans on the data:

$$\phi_{nk} \propto \exp\left(-\frac{1}{2}||x_n - \mu_k||\right)$$

3.1 Identifying the Correct Clustering

In the first experiment, we generate data from 4 Gaussian distributions, 5 % of the samples of the two clusters at the bottom and the top are labeled, for the two clusters in the middle we label 10 % of all the samples with the same label different from the previous two (Figure 2).

We apply a classical Dirichlet Process Gaussian Mixture Model, and our Semi-Supervised Dirichlet Process Gaussian Mixture Model on this dataset. We set the truncation level of the number of clusters to $T = 4$ and then to $T = 10$, with the same value for the concentration parameter $\eta = 1$ of the Dirichlet Process. We depict the clustering process across iterations for both the Dirichlet Process Gaussian Mixture Model and our Semi-Supervised Dirichlet Process Gaussian Mixture Model with the truncation level of $T = 4$ in Figure 3 and for $T = 10$ in Figure 4.

In both cases the semi-supervised DPGMM identifies the correct number of clusters ($K = 3$) given the partial labels, unlike the classical DPGMM where the number of clusters identified depends on the concentration parameter and the initialization. The intuition gained from this experiment is that the partial labels help the Dirichlet process to squeeze out unnecessary clusters to explain the data, even if the concentration parameter is high (tendency to produce high number of clusters).

3.2 Identifying the Correct View

In the second experiment, in the same fashion, we generate 400 samples from 4 Gaussian distributions as shown in Figure 6. In this case the clusters are arranged in such a way that if $K = 2$, two possible views of the clustering are possible. To identify the correct clustering some information about the view is needed. We suppose that we have two sets of partial labels, and as Figure 7 shows these partial labels encode which view is the correct one in both cases.

We apply the semi-supervised DPGMM in both cases (Figure 5), we set the truncation level of the number of cluster $T = 3$. We notice that in both cases starting from a KMeans initialization the semi-supervised DPGMM adapts to reach the clustering that respects the view imposed by the partial labels.

3.3 Improving the Clustering Accuracy

In the last experiment, we apply the semi-supervised DPGMM on the classical UCI datasets (wine, digits, iris, glass, yeast). We consider that the true number of clusters is unknown. We set the truncation level to $T = 10$ for wine, and iris, and $T = 20$ for digits, glass, and yeast. The evaluation metric used to evaluate the clustering is the clustering accuracy [15] defined as :

$$\text{ACC} = \max_{m \in \mathcal{M}} \frac{\sum_{n=1}^{N} \mathbb{1}\left[l_n = m(c_n)\right]}{N}$$

where c_n is the cluster assignment, l_n the true label and $\mathcal{M}$ the set of all possible one-to-one mappings. we vary the percentage of partially labeled samples from 0% (no labels known, fully unsupervised) to 50%, by a step of 10% . On each dataset we report the best accuracy over 10 reruns of the model, and we plot the evolution of the accuracy for each dataset as a function of the percentage of partial labels (Figure 8). We notice that the clustering accuracy improves as we add labels. For the iris dataset, for example, we notice that the introduction of 20% of the labels improves the accuracy from 72% to 98%. In table 1 we report the clustering accuracy for the classical DPGMM, the semi-supervised DPGMM at 20% of partial labels and at 50% of partial labels.

Datasets	DPGMM (0%)	SS-DPGMM (20%)	SS-DPGMM (50%)
Iris	.72	.98	.98
Wine	.49	.58	.81
Glass	.52	.5	.71
Yeast	.37	.41	.70
Digits	.61	.68	.79

Table 1: Accuracy score of the Semi-Supervised Dirichlet Process Gaussian Mixture Model (SS-DPGMM) for multiple percentages of partial labels.

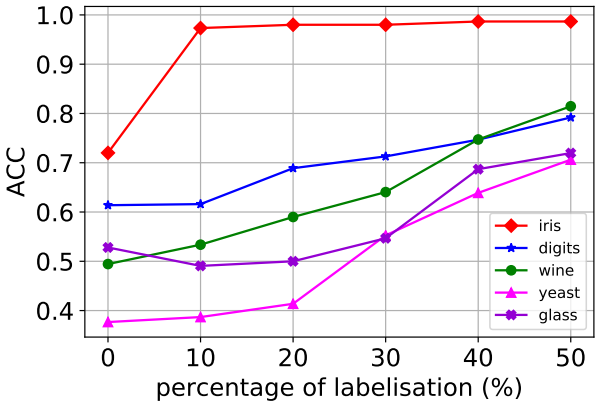

Figure 8: Accuracy of the semi-supervised DPGMM on the UCI datasets. The percentage of partial labels varies from 0 to 50 % of the true labels.

3.4 Comparison with HMRF-KMeans

In this section, we compare our approach to the HMRF-KMeans semi-supervised clustering algorithm [3] which is a strong baseline for the semi-supervised clustering task. Similarly to the previous experiment we evaluate the HMRF-KMeans on the classical UCI datasets. We fix the number of clusters to the true number of classes in the dataset and we report the clustering accuracy for different percentages of partial labels. Figure 9 shows the evolution of the clustering accuracy for each dataset.

Our approach clearly outperforms the HMRF-KMeans in terms of clustering accuracy. For example, we can see that for 50% labelisation the clustering accuracy for all datasets is between [0.5, 0.7], while for our method the accuracy is above 0.7. Furthermore, HMRF-KMeans uses Iterated Conditional modes [5] during the E-Step. This algorithm increases the complexity and therefore the time of execution for each iteration of the EM, unlike our approach, which is based on the classical Variational EM. Our method can identify new separate unlabeled clusters, and thus learn the number of clusters automatically thanks to the Dirichlet Process prior. Unlike the HMRF-Kmeans where the true number of clusters has to be set in advance.

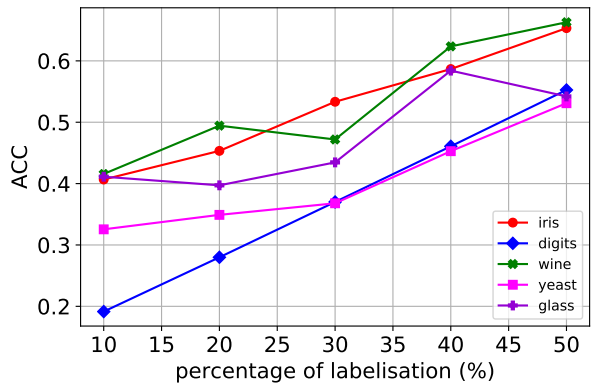

Figure 9: Accuracy of the HMRF-KMeans algorithm on the UCI datasets. The percentage of partial labels varies from 10 to 50 % of the true labels.

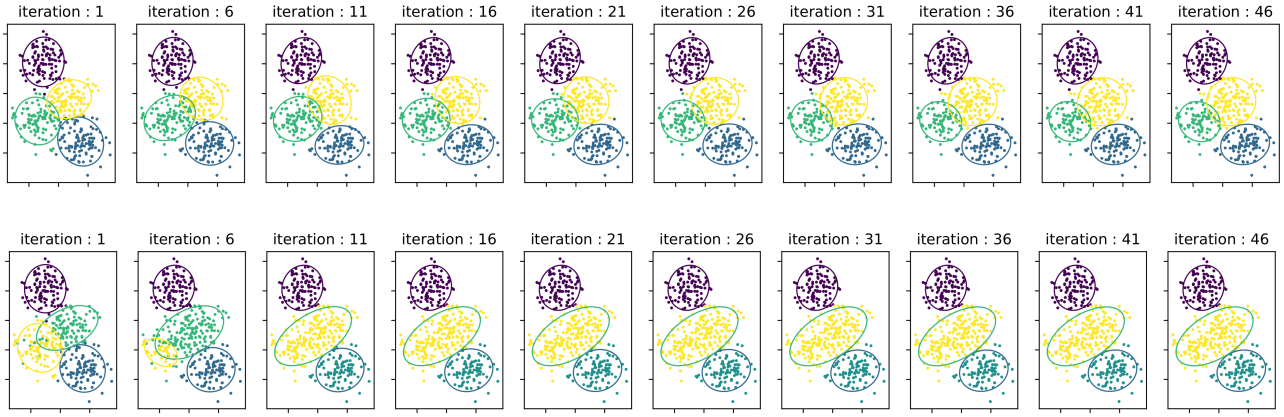

Figure 3: Visualization of the clustering process across iterations of the classical DPGMM $T = 4$ (top), and the semi-supervised DPGMM $T = 4$ (bottom).

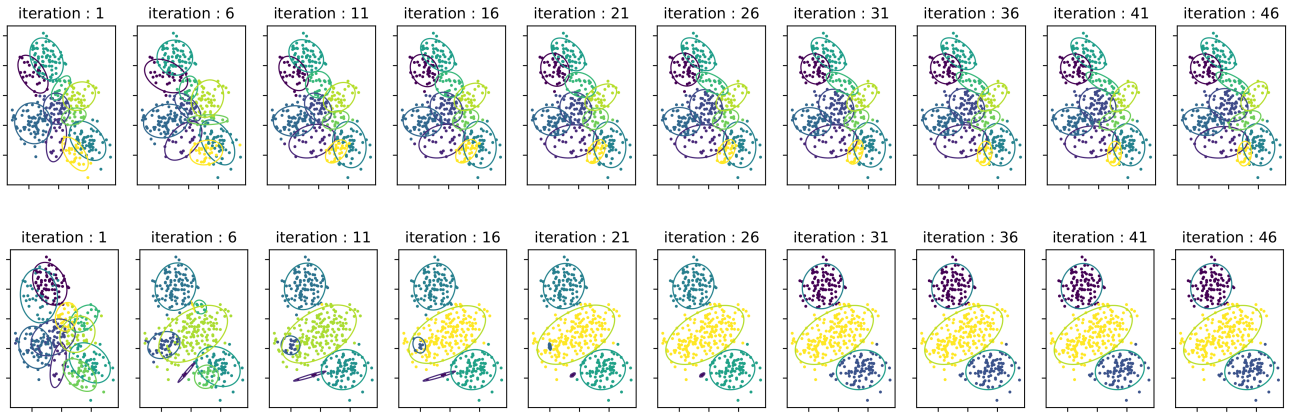

Figure 4: Visualization of the clustering process across iterations of the classical DPGMM $T = 10$ (top), and the semi-supervised DPGMM $T = 10$ (bottom).

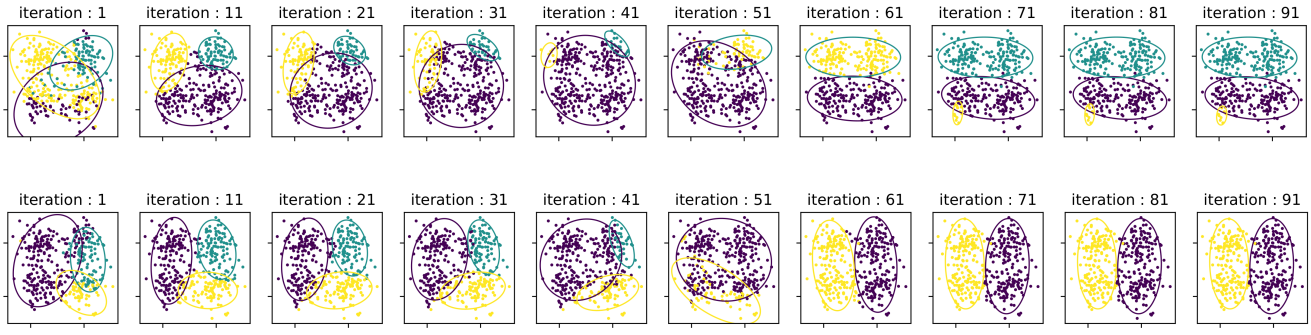

Figure 5: The clustering process across iterations in the case of partial labels corresponding to the horizontal clustering (top), and vertical clustering (bottom).

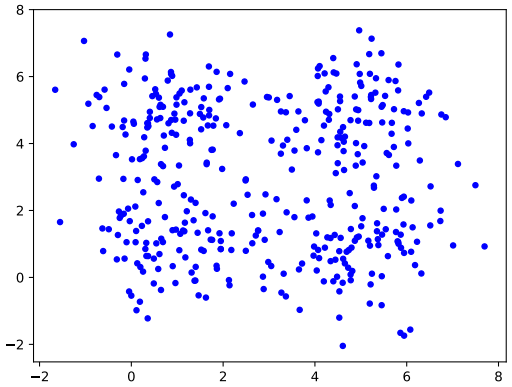

Figure 6: Samples from 4 Gaussian distributions, if $K = 2$ two views of clustering using a linear boundary are possible: horizontal (two top clusters separated from the bottom clusters) and vertical (the two clusters on the left separated from the clusters on the right).

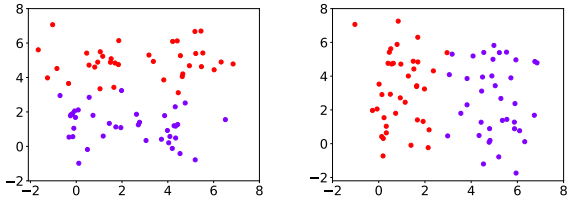

Figure 7: Partial labels corresponding to the horizontal clustering (left), partial labels corresponding to the vertical clustering (right).

4 Conclusion

In this paper, we introduced a new method to perform semi-supervised clustering with Bayesian finite and infinite mixture models. This approach allows the introduction of prior knowledge in the form of partial labels set by an expert to guide the clustering process towards the correct solution. We have shown that the model can identify the correct view of the clustering and the correct number of clusters, we also demonstrated that by introducing a small fraction of partial labels we can improve the overall accuracy of a classical mixture model. Our approach is general, and can easily be applied to other types of mixture models like the categorical mixture model, latent Dirichlet allocation or topic models. In future work, we will explore how we can extend this approach to more complex probabilistic graphical models. We will also investigate stochastic variational inference [9] in order to apply the approach to large scale datasets.

A Semi-Supervised Dirichlet Process Gaussian Mixture Model

In this section, we develop the mean field update equations in the case of the Dirichlet Process Gaussian Mixture Model. In what follows we adopt the formulation presented in [11].

The generative process of the model is the following:

$$\begin{aligned}
\forall k \quad \beta_k &\sim \text{Beta}(1, \eta) \\
\pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \\
\mu_k | \Lambda_k &\sim \mathcal{N}(\cdot | m_0, (\kappa_0 \Lambda_k)^{-1}) \\
\Lambda_k &\sim \mathcal{W}(\cdot; L_0, \nu_0) \\
\mathbf{z}_{1:N} | \pi &\sim p_{\mathbf{z}_{1:N}}(\cdot | \pi) \\
\mathbf{x}_n | \mathbf{z}_n = k, \mu, \Lambda &\sim \mathcal{N}(\cdot | \mu_k, \Lambda_k^{-1})
\end{aligned}$$

Where $\mathcal{W}$ is the Wishart distribution. By substituting in the mean field update equations, we have $\forall k \in \{1, .., T\}$:

$$\begin{aligned}
q(\beta_k) &= \text{Beta}(\beta_k; \gamma_{1,k}, \gamma_{2,k}) \\
q(\mu_k | \Lambda_k) &= \mathcal{N}(\mu_k; m_k, (\kappa_k \Lambda_k)^{-1}) \\
q(\Lambda_k) &= \mathcal{W}(\Lambda_k; L_k, \nu_k) \\
q(\mathbf{z}_n) &= \text{Cat}(\mathbf{z}_n; \phi_n)
\end{aligned}$$

The variational parameters have the following fixed point equations:

$$\begin{aligned}
\gamma_{1,k} &= 1 + \sum_{n=1}^{N} \phi_{nk} \quad \gamma_{2,k} = \eta + \sum_{n=1}^{N} \sum_{l=k+1}^{T} \phi_{nl} \\
\kappa_k &= \kappa_0 + \sum_{n=1}^{N} \phi_{nk} \quad \nu_k = \nu_0 + \sum_{n=1}^{N} \phi_{nk} + 1 \\
m_k &= \frac{\kappa_0 m_0 + \sum_{n=1}^{N} \phi_{nk} \mathbf{x}_n}{\kappa_k} \\
L_k^{-1} &= L_0^{-1} + \kappa_0 (m_k - m_0)(m_k - m_0)^T \\
&+ \sum_{n=1}^{N} \phi_{nk} (\mathbf{x}_n - m_k)(\mathbf{x}_n - m_k)^T
\end{aligned}$$

$$\begin{aligned}
\log \phi_{nk} &= -\frac{1}{2} \left[\frac{d}{\kappa_k} + \nu_k (\mathbf{x}_n - m_k)^T L_k (\mathbf{x}_n - m_k) \right] \\
&+ \frac{1}{2} \left[\sum_{i=1}^{d} \psi \left(\frac{\nu_k + 1 - i}{2} \right) + \log |L_k| \right] \\
&+ \psi(\gamma_{1,k}) - \psi(\gamma_{1,k} + \gamma_{2,k}) \\
&+ \sum_{l=1}^{k-1} [\psi(\gamma_{2,l}) - \psi(\gamma_{1,l} + \gamma_{2,l})] \\
&- \lambda \sum_{m \in \mathcal{N}_n} \sum_{l=1}^{T} \phi_{ml} \mathcal{V}(k, l) + \text{const}
\end{aligned}$$

where ψ is the digamma function and as defined in (4) the expression of $\mathcal{V}$ is :

$$\mathcal{V}(k, l) = \mathbb{D}_{KL}^{sym} \left[q(\mu_k, \Lambda_k) || q(\mu_l, \Lambda_l) \right]$$

which we can compute in close form as a function of the variational parameters of $q(\mu_k, \Lambda_k)$ and $q(\mu_l, \Lambda_l)$, using the standard formulas for the kullback-leibler divergences between two Gaussian distributions and between two Wishart distributions:

$$
\begin{aligned}
\mathcal{V}(k,l) &= d(\frac{\kappa_l}{\kappa_k}+\frac{\kappa_k}{\kappa_l}-2)+\frac{1}{2}(\nu_k-\nu_l)(\log(|L_k|)-\log(|L_k|)) \\
&+ \operatorname{Tr}\left((\nu_k L_k+\nu_l L_l)(m_k-m_l)(m_k-m_l)^T\right) \\
&+ \frac{1}{2}(\nu_k-\nu_l)\left[\sum_{i=1}^{d}\psi\left(\frac{\nu_k+1-i}{2}\right)\right] \\
&+ \frac{1}{2}(\nu_l-\nu_k)\left[\sum_{i=1}^{d}\psi\left(\frac{\nu_l+1-i}{2}\right)\right] \\
&+ \frac{1}{2}\operatorname{Tr}\left(\nu_k L_l^{-1}L_k+\nu_l L_k^{-1}L_l\right)-\frac{d}{2}(\nu_k+\nu_l)
\end{aligned}
$$

B The Normalizing Constant Γ

By definition the normalizing constant Γ can be written as :

$$
\begin{aligned}
\Gamma &= \sum_{\mathbf{z}_{1:N}}\prod_{n=1}^{N}\pi_{\mathbf{z}_n}^{\mathbb{1}[\mathcal{N}_n=\emptyset]}\prod_{n\sim m}e^{-\lambda\mathcal{V}(\mathbf{z}_n,\mathbf{z}_m)} \\
&= \sum_{\substack{\mathbf{z}_n\forall n \\ \text{s.t}\,\mathcal{N}_n\neq\emptyset}}\left[\sum_{\substack{\mathbf{z}_n\forall n \\ \text{s.t}\,\mathcal{N}_n=\emptyset}}\prod_{n=1}^{N}\pi_{\mathbf{z}_n}^{\mathbb{1}[\mathcal{N}_n=\emptyset]}\right]\prod_{n\sim m}e^{-\lambda\mathcal{V}(\mathbf{z}_n,\mathbf{z}_m)}
\end{aligned}
$$

Let's denote by $\mathcal{S}$ the set containing all points with empty neighborhoods :

$$\mathcal{S}=\{n \text{ s.t } \mathcal{N}_n=\emptyset\}$$

We have:

$$
\begin{aligned}
\sum_{\substack{\mathbf{z}_n\forall n \\ \text{s.t}\,\mathcal{N}_n=\emptyset}}\prod_{n=1}^{N}\pi_{\mathbf{z}_n}^{\mathbb{1}[\mathcal{N}_n=\emptyset]} &= \sum_{\substack{\mathbf{z}_n \\ \forall n\in\mathcal{S}}}\prod_{n\in\mathcal{S}}\pi_{\mathbf{z}_n} \\
&= \prod_{n\in\mathcal{S}}\sum_{\mathbf{z}_n}\pi_{\mathbf{z}_n} \\
&= 1
\end{aligned}
$$

Thus:

$$\Gamma=\sum_{\mathbf{z}_n,\forall n\text{ s.t }\mathcal{N}_n\neq\emptyset}\prod_{n\sim m}e^{-\lambda\mathcal{V}(\mathbf{z}_n,\mathbf{z}_m)}$$

REFERENCES

[1] Eric Bair, 'Semi-supervised clustering methods', *Wiley Interdisciplinary Reviews: Computational Statistics*, **5**(5), 349–361, (2013).

[2] Sugato Basu, Arindam Banerjee, and Raymond J Mooney, 'Active semi-supervision for pairwise constrained clustering', in *Proceedings of the 2004 SIAM international conference on data mining*, pp. 333–344. SIAM, (2004).

[3] 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*, pp. 59–68. ACM, (2004).

[4] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux, '11 label propagation and quadratic criterion', (2006).

[5] Julian Besag, 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(3), 259–279, (1986).

[6] David M Blei, Michael I Jordan, et al., 'Variational inference for dirichlet process mixtures', *Bayesian analysis*, **1**(1), 121–143, (2006).

[7] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, 'Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]', *IEEE Transactions on Neural Networks*, **20**(3), 542–542, (2009).

[8] John M Hammersley and Peter Clifford, 'Markov fields on finite graphs and lattices', *Unpublished manuscript*, **46**, (1971).

[9] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley, 'Stochastic variational inference', *The Journal of Machine Learning Research*, **14**(1), 1303–1347, (2013).

[10] Thorsten Joachims, 'Transductive inference for text classification using support vector machines', in *Icml*, volume 99, pp. 200–209, (1999).

[11] Kevin P Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[12] Jayaram Sethuraman, 'A constructive definition of dirichlet priors', *Statistica sinica*, 639–650, (1994).

[13] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al., 'Constrained k-means clustering with background knowledge', in *Icml*, volume 1, pp. 577–584, (2001).

[14] Martin J Wainwright, Michael I Jordan, et al., 'Graphical models, exponential families, and variational inference', *Foundations and Trends® in Machine Learning*, **1**(1–2), 1–305, (2008).

[15] Junyuan Xie, Ross Girshick, and Ali Farhadi, 'Unsupervised deep embedding for clustering analysis', in *International conference on machine learning*, pp. 478–487, (2016).

[16] Yongyue Zhang, Michael Brady, and Stephen Smith, 'Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm', *IEEE transactions on medical imaging*, **20**(1), 45–57, (2001).

[17] Xiaojin Zhu and Zoubin Ghahramani, 'Learning from labeled and unlabeled data with label propagation', (2002).