Łukasz Rajkowski 2 March 2021

Maximal a Posteriori Partition in Nonparametric Bayesian Mixture Models with applications to Clustering Problems

Self-report

1 General description of the research problem

Let us start with an explanation of the term *clustering*. In data analysis it is an operation of distinguishing heterogeneous subgroups in the given dataset. In other words, clustering is the task of finding a proper *partition* of the dataset. In two-dimensional setting clustering can be done even by a simple visual assessment. This becomes impossible as the dimensionality of the problem increases – hence the need of an algorithmic approach.

The approach of the dissertation is *Bayesian*, meaning that we put a probability distribution (called *prior*) on the space of all possible partitions. For every partition we assign (in a natural way) a conditional probability distribution of data given this partition. This allows us, using Bayes Rule, to inverse the conditionality and compute the conditional probability distribution on the space of partitions given the data (this is the *posterior*) that encapsulates all the information about the partition structure that we can (in our model) infer from the data. This is easily said, but the computation of the exact probability weight is practically impossible as it requires summing over the whole space of partitions, which is intractable. Thanks to Markov Chain Monte Carlo algorithms, it is however possible to sample from the approximation of the posterior distribution, hence this approach to clustering is of interest also for practitioners. We also restrict our attention to priors on the partition space that are obtained from some probability distribution on the infinitely dimensional space of label probabilities. We call such models *Bayesian Mixture Models* (BMMs).

If a concrete estimate of the partition structure is needed (which is a standard case in applications), a natural solution is to pick the partition that maximises the posterior probability (i.e. the most probable partition given the data, and the model of course). Such partition is called *the Maximum a Posteriori*, or MAP, partition. This is also a computationally convenient choice since in order to find the maximiser it is enough to know the posterior probability weights up to the problematic norming constant, and such quantity is easily obtained via the product rule. Still the partition space is too large for exhaustive search, but at least we can easily compare two different partitions in terms of this 'posterior score'.

In the dissertation, we limit our attention to conjugate exponential families as a mechanism of generating the data within clusters. Conjugate exponential family is a general and important class of distributions widely used in Bayesian analysis. A popular example is the situation in which the data within every cluster has a multivariate normal distribution in which the mean itself is also sampled from multivariate normal distribution; we will call it the Normal-Normal case. We also put some special attention to a model in which the data within clusters is normally distributed, but not only the cluster mean is random and distributed normally, but also the covariance matrix is random and distributed according to the Inverse-Wishart distribution. We call this situation the Normal-Inverse-Wishart case.

The MAP partition is the main object of our interest. We prove that in the conjugate exponential BMM the clusters of the MAP partition must be separated by the contour surfaces of linear functionals of the sufficient statistics. In other words, if we use the sufficients statistics instead of the original data, the clusters become linearly separated, i.e. their convex hulls are disjoint. In this sense the clusters in the MAP partition can be thought as being defined by a decent partition of the observation space (where 'being defined' means that the data placed in the same chunk of the observation space are clustered together and 'decent' means that the chunks are counterimages of convex polytopes under the sufficient statistic). Of course, the partition of the observation space that defines the MAP clustering can change as the number of observations increases. Nevertheless it seems interesting to analyse the posterior probability of clusterings that are defined by a fixed partition of the observation space (we call such clusterings *induced clusterings*). We derive the formula for the asymptotic limit (up to a constant) of the logarithm of the posterior probability of an induced partition in conjugate exponential BMM, when the data is an independent sample from some probability distribution P, called the input distribution. Interestingly, the limit does not depend on the prior probability on mixture weights, provided the latter has a full support on an infinitely dimensional simplex. The aforementioned asymptotic limit is a function of the partition of the observation space – we call it the Δ function, since it is a difference of two functions that increase their values whenever two chunks of the partitions are merged and therefore the maximisation of this function represents a trade-off between two tendencies: fine partitions adjust well to the data but at the same time they are penalized by the prior. A natural idea there is that perhaps the MAP clusterings are somehow bound to the maximisers of the Δ function. This line of research was pursued in Rajkowski (2019), where the positive result was proved for a very specific example of an conjugate exponential BMM, namely the Normal-Normal BMM (we later call a Normal-Normal BMM) with the so called *Chinese Restaurant* prior, explained later in this report. These findings are also presented in details in the dissertation.

The fixed covariance model clearly imposes severe limitations on the covariance structure within clusters, rarely met in the real world situations. Models that differences between the covariance structures of the clusters should perform better when clusters do have different covariance structures. We attempt to deal with this in the Normal-Inverse-Wishart model. At the same time, we observed some undesired behaviour of the Δ_P function for this model. For example when the input distribution is uniform on a segment, in which case every partition into subsegments gives the same Δ_P score. This is why we also consider an adjusted Normal-Inverse-Wishart model, where the concentration parameter of the prior on the covariance structure is increasing linearly with the number of observations. It turns out that with this model, we can rewrite some of the results from Rajkowski (2019). Finally, in this case as a limit we obtain a family of Δ_P functions that depend on the linear coefficient in the concentration parameter. We can translate this Δ_P functions to their empirical counterparts and hence obtain a convenient family of score function the measure the performance of data clustering. This, or rather its empirical equivalent, can be used for scoring candidates for partitions proposed by some more ad-hoc methods, like the k-means. This approach is investigated in numerical simulations, presented towards the end of the dissertation.

2 Mathematical setting

2.1 General Framework for BMMs

We start with the description of Bayesian Mixture Models. Let $\Theta \subset \mathbb{R}^p$ be the parameter space for a single cluster distributions and $\{G_\theta \colon \theta \in \Theta\}$ be a family of probability measures on the observation space $\mathbb{R}^d$ and assume that G_θ has a density g_θ with respect to the Lebesgue measure. Those are the *component* measures, responsible for randomness within clusters. Consider a prior distribution ϑ on Θ (we will call it the *base* measure, defining how the parameters of the components are spread). Let π be a prior probability distribution on the m-dimensional simplex $\triangle^m = \{\boldsymbol{p} = (p_i)_{i=1}^m \colon \sum_{i=1}^m p_i = 1 \text{ and } p_i \geq 0 \text{ for } i \leq m\}$ (where $m \in \mathbb{N} \cup \{\infty\}$). The observations $x_1, \ldots, x_n \in \mathbb{R}^d$ are modelled by

$$\begin{array}{rcl} \boldsymbol{p} = (p_i)_{i=1}^m & \sim & \pi \\ \boldsymbol{\theta} = (\theta_i)_{i=1}^m & \overset{\text{iid}}{\sim} & \vartheta \\ \boldsymbol{x} = (x_1, \ldots, x_n) \,|\, \boldsymbol{p}, \boldsymbol{\theta} & \overset{\text{iid}}{\sim} & \sum_{i=1}^m p_i G_{\theta_i}. \end{array} \tag{1}$$

This is a *Bayesian Mixture Model.* If $m < \infty$ we call the model *finite*, otherwise it is (obviously) *infinite*. In this dissertation we concentrate on the infinite case.

The focus of this dissertation is applying Bayesian Mixture Models to detect clusters within data. Indeed, formula (1) can be used to model data clustering; clusters are defined by deciding which distribution G_{θ_i} generated a given data point. This can be easily formalized as a probability distribution $\mathcal{P}_{\pi,n}$ on the space of all partitions of $[n] := \{1, 2, \ldots, n\}$. We can now formulate (1) as follows: firstly we generate the partition of observations into clusters according to $\mathcal{P}_{\pi,n}$, and then for each cluster we sample actual observations from the relevant marginal distribution on the data. To formalise this description succinctly, we introduce some additional notation. If $\boldsymbol{x} = (x_i)_{i=1}^n$ is a sequence and $I \subseteq [n]$, then $\boldsymbol{x}_I = (x_i)_{i \in I}$ is a subsequence of $\boldsymbol{x}$ consisting of the terms at coordinates belonging to I. The distribution $G_{\vartheta,k}$ ($k \in \mathbb{N}$) is the marginal distribution of the k-tuple whose coordinates are, conditionally on $\theta \sim \vartheta$, independently and identically distributed by G_θ. More specifically, for $\theta \sim \vartheta$, $k \in \mathbb{N}$ and $\boldsymbol{u} = (u_1, \ldots, u_k) \,|\, \theta \overset{\text{iid}}{\sim} G_\theta$, we denote by $G_{\vartheta,k}$ the marginal distribution of $\boldsymbol{u}$. Its density is given by

$$g_{\vartheta,k}(u_1, \ldots, u_k) := \int_\Theta \prod_{i=1}^{k} g_\theta(u_i) \mathrm{d}\vartheta(\theta). \tag{2}$$

Now, (1) is equivalent to

$$\begin{array}{rcl} \mathcal{I} & \sim & \mathcal{P}_{\pi,n} \\ \boldsymbol{x}_I := (x_i)_{i \in I} \,|\, \mathcal{I} & \sim & G_{\vartheta,|I|} \quad \text{for all } I \in \mathcal{I} \end{array} \tag{3}$$

Using the within cluster conditional independence, we can write the density of $\boldsymbol{x}$ conditionally on $\mathcal{I}$:

$$g_{\vartheta,n}(\boldsymbol{x} \,|\, \mathcal{I}) := \prod_{I \in \mathcal{I}} g_{\vartheta,|I|}(\boldsymbol{x}_I). \tag{4}$$

Finally, for further convenience, let

$$Q(\boldsymbol{x}, \mathcal{I}) = \mathcal{P}_{\pi,n}(\mathcal{I}) \cdot g_{\vartheta,n}(\boldsymbol{x} \,|\, \mathcal{I}) \tag{5}$$

be the joint density of the partition and the observation. By Bayes rule, the expression in (5) is also proportional to the posterior probability $\mathcal{P}_{\pi,n}(\mathcal{I} \,|\, \boldsymbol{x})$ of the partition $\mathcal{I}$ given the observation $\boldsymbol{x}$.

The expression in Formula (5) is proportional to the posterior distribution on the space of partitions. Therefore, the maximiser of this expression gives the Maximum A Posteriori clustering and this is what we use as an estimator of the clustering structure.

Definition 2.1. The *Maximum A Posteriori* (MAP) partition of $[n]$ given $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ in a given Bayesian Mixture Model of the form (3) is any partition $\hat{\mathcal{I}}$ of $[n]$ that maximises $Q(\boldsymbol{x}, \mathcal{I})$ given by (5). In other words, the set of the MAP partitions is given by $\operatorname{argmax}_{\mathcal{I}} Q(\boldsymbol{x}, \mathcal{I})$.

A classical choice for the partition prior $\mathcal{P}_{\pi,n}$ is so called *Chinese Restaurant Process.* The construction goes as follows: imagine that elements of $[n]$ are the clients waiting in front of a Chinese Restaurant, in which there is potentially infinitely many tables. Customer 1 chooses any table she wants. Customer 2 chooses another table with probability proportional to α or joins Customer 1 with probability proportional to 1; thus those probabilities are $\frac{\alpha}{\alpha+1}$ and $\frac{1}{\alpha+1}$ respectively. In general, the n-th customer chooses an empty table with probability proportional to α or joins a nonempty table with probability proportional to the number of other customers sitting there. This description is readily transformed into the following probability function on the space of all partitions of $[n]$:

$$\mathcal{P}_{\pi,n}(\mathcal{I}) = \frac{\alpha^{|\mathcal{I}|}}{\alpha^{(n)}} \prod_{I \in \mathcal{I}} (|I| - 1)!, \tag{6}$$

where $\alpha^{(n)} = \alpha(\alpha+1)\dots(\alpha+n-1)$.

Now we briefly present a natural and computationally convenient candidates for distributions ν and G_θ (the base and the component measures), namely conjugate exponential families. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the observation space and let $\Theta \subseteq \mathbb{R}^p$ be the parameter space. A family of distributions $\{G_\theta \colon \theta \in \Theta\}$ on $\mathcal{X}$ is called *p-dimensional exponential family* if for every θ the probability G_θ has the following density with respect to the Lebesgue measure:

$$g_\theta(x) = h(x) \cdot \exp\left\{T(x)^\top \eta(\theta) - \mathsf{B}(\theta)\right\}, \tag{7}$$

where $T \colon \mathcal{X} \to \mathbb{R}^p$ is a p-dimensional statistic (called *natural sufficient statistic*) and $h \colon \mathcal{X} \to \mathbb{R}$, $\mathsf{B} \colon \Theta \to \mathbb{R}$ and $\eta \colon \Theta \to \mathbb{R}^p$ are some functions.

If we let the model be indexed by $\eta = \eta(\theta)$ rather than θ we obtain *the canonical p-parameter exponential family generated by T and h*, in which the density of $G'_\eta = G_\theta$ is given by

$$g'_\eta(x) = h(x) \cdot \exp\left\{T(x)^\top \eta - \mathsf{A}(\eta)\right\}, \tag{8}$$

where

$$\mathsf{A}(\eta) = \log \int_{\mathcal{X}} h(x) \cdot \exp\left\{T(x)^\top \eta\right\} \mathrm{d}x \tag{9}$$

is called *the log-partition function.* In this case the set

$$\mathcal{E} = \{\eta \in \mathbb{R}^p \colon \mathsf{A}(\eta) < \infty\} \tag{10}$$

is called *the natural parameter space.* If the natural parameter space is a nonempty open subset of $\mathbb{R}^p$, we say that the canonical exponential family is *regular.* Moreover we will use the term *regular* for an exponential family $\{g_\theta \colon \theta \in \Theta\}$ (where g_θ is given by (7)) when the corresponding canonical form is regular and $\theta \colon \Theta \to \mathcal{E}$ is a bijection.

Now we introduce a *conjugate exponential family*, i.e. an exponential family of distributions such that if we consider a Bayesian model in which the prior distribution on the parameter θ comes from this family and the likelihood is given by (7), then the posterior distribution $\theta \,|\, \boldsymbol{x}$ also belongs to this family.

Suppose that in (7) we can write $\mathsf{B}(\theta)$ as $\mathsf{B}(\theta) = \boldsymbol{a}^\top \boldsymbol{\mathsf{B}}(\theta)$ where $\boldsymbol{a} \in \mathbb{R}^q$ and $\boldsymbol{\mathsf{B}}(\theta) = [\mathsf{B}_1(\theta), \dots, \mathsf{B}_q(\theta)]^\top$. Consider a canonical exponential family on Θ, where the densities are given by

$$\gamma_{\tau,\zeta}(\theta) := \psi(\theta) \cdot \exp\left\{[\eta(\theta)^\top, -\boldsymbol{\mathsf{B}}(\theta)^\top] \begin{bmatrix} \tau \\ \zeta \end{bmatrix} - \mathsf{C}(\tau,\zeta)\right\}, \tag{11}$$

where $\tau \in \mathbb{R}^p$ and $\zeta \in \mathbb{R}^q$ are the hyperparameters and $\mathsf{C}(\tau,\zeta)$ is the log-partition function. It follows that if $\theta \sim \vartheta$, where ϑ has density γ_{τ_0,ζ_0} and $\boldsymbol{x} = (x_1, \dots, x_k) \,|\, \theta \overset{\text{iid}}{\sim} g_\theta$ then $\theta \,|\, \boldsymbol{x} \sim \gamma_{\tau_{\boldsymbol{x}},\zeta_k}$, where $\tau_{\boldsymbol{x}} := \tau_0 + \sum_{i=1}^k T(x_i)$ and $\zeta_k := \zeta_0 + k\boldsymbol{a}$. Moreover, the marginal density of $\boldsymbol{x}$ is

$$\boldsymbol{x} \sim g_{\vartheta,k}(\boldsymbol{x}) = \prod_{i=1}^k h(x_i) \cdot \exp\left\{\mathsf{C}(\tau_{\boldsymbol{x}},\zeta_k) - \mathsf{C}(\tau_0,\zeta_0)\right\} \tag{12}$$

Convexity assumption. It is a well known property of the exponential families that $\boldsymbol{a}^\top \mathbf{B}\big(\theta(\eta)\big) = \mathsf{B}\big(\theta(\eta)\big) = \mathsf{A}(\eta)$ is a convex function on $\mathcal{E}$. In some of our results we will assume that also

$$\text{for any } (\tau_0, \zeta_0) \in \Omega \text{ the function } {\zeta_0}^\top \mathbf{B}\big(\theta(\eta)\big) \text{ is a convex function on } \mathcal{E}. \tag{13}$$

We call this assumption a *convexity assumption*; it is satisfied by all multivariate conjugate Normal models presented in the dissertation.

Definition 2.2. *Canonical Exponential Family Bayesian Mixture Model* is a Bayesian Mixture Model in which the component density is given by (7) and the base density is (11) for some $(\tau_0, \zeta_0) \in \Omega$.

2.2 Example: Conjugate Normal Families

As an example of conjugate exponential family that is commonly used in practice (in the context of mixture models) we consider *Normal* Conjugate Families in which the component distributions G_θ are multivariate Normal. This corresponds to the data being normally distributed within clusters, which is a rather standard assumption.

Normal-Normal (NN). Here the component covariance matrix is assumed to be known a priori; the component mean is unknown and this is the parameter on which the prior distribution is set, i.e. $\theta = \mu$, $\Theta = \mathbb{R}^d$ and $x \mid \mu \sim \mathcal{N}(\mu, \Sigma_0)$, where Σ_0 is known. The base measure is

$$\mu \quad \sim \quad \mathcal{N}(\mu_0, \Psi_0). \tag{14}$$

Normal-Inverse-Wishart (NIW). In this case both the mean and the covariance matrix are unknown. The parameter space is therefore equal to $\Theta = \mathbb{R}^d \times \mathcal{S}_+^d$, where $\mathcal{S}_+^d$ is the space of all positive definite, $d \times d$ matrices, that can serve as convariance structures. For $\theta = (\mu, \Lambda) \in \Theta$ the component distribution is $x \mid \theta \sim \mathcal{N}(\mu, \Lambda)$ and the base measure ϑ on (μ, Λ) is defined by the following conditional structure

$$\begin{aligned} \Lambda \quad &\sim \quad \mathcal{W}^{-1}(\nu_0 + d + 1, \nu_0 \Sigma_0) \\ \mu \mid \Lambda \quad &\sim \quad \mathcal{N}(\mu_0, \Lambda/\kappa_0), \end{aligned} \tag{15}$$

where $\mathcal{W}^{-1}$ is the Inverse-Wishart distribution.

3 Statement of the results

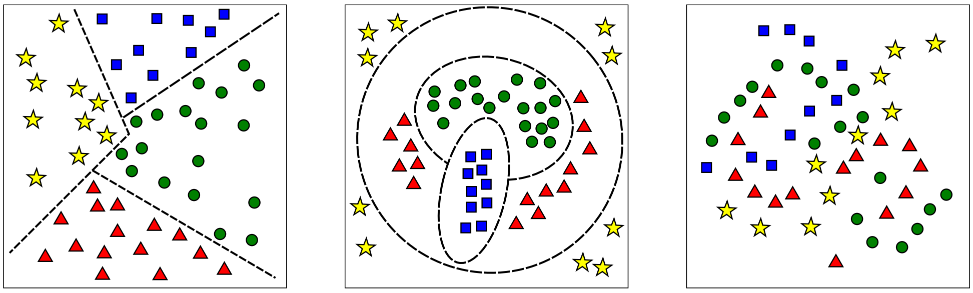
3.1 Geometric separability

The first important result of the dissertation concerns the separation of clusters in the MAP partition. In Rajkowski (2019, Proposition 1) it was proved that for the Gaussian fixed covariance BMM model (with the Chinese Restaurant prior on the space of partitions), the convex hulls of the clusters in the MAP partition are disjoint. In other words, every two clusters are separated by a hyperplane or linear affine subspace. Theorem 3.3 generalises that result to the conjugate exponential BMMs and shows how the separability property of clusters relates to the sufficient statistic $T(x)$ in the conjugate exponential family. More precisely, in the general case the separation surfaces are the contour lines of linear functionals of the sufficient statistic.

We start this section by defining what we mean by T-linear separation of clusters.

Definition 3.1. Let $\mathcal{Z}$ be a family of subsets of $\mathbb{R}^d$ and $\mathcal{L}$ a family of real functions on $\mathbb{R}^d$. We say that $\mathcal{Z}$ *is separated by* $\mathcal{L}$ if for every $A, B \in \mathcal{Z}$, $A \neq B$, there exists $L_{A,B} \in \mathcal{L}$ such that $L_{A,B}(x) \geq 0$ and $L_{A,B}(y) < 0$ for all $x \in A, y \in B$. Moreover, if $\mathcal{L} = \{\boldsymbol{a}^\top T(x) + b \colon \boldsymbol{a} \in \mathbb{R}^p, b \in \mathbb{R}\}$ for some function $T \colon \mathbb{R}^d \to \mathbb{R}^p$, we say that $\mathcal{Z}$ *is T-linearly separated.* If $T(x) = x$, we use the term *linear separability* for short.

Note 3.2. If a family $\mathcal{Z}$ of subsets of $\mathbb{R}^d$ is linearly separable, then every pair of elements of $\mathcal{Z}$ is separated (in standard, geometric sense) by a hyperplane.



(a) This family is linearly separable.

(b) This family is quadratically ($T(x) = [\mathrm{diag}(xx^\top), \mathrm{low}(xx^\top), x]$) separable. It is not linearly separable.

(c) This family is not quadratically separable.

Figure 1: Illustration of the different types of separability. The family $\mathcal{Z}$ in each picture consists of four sets: stars, sqares, triangles and circles (distinguished also by color).

Theorem 3.3. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be pairwise distinct and let $\hat{\mathcal{I}}$ be the MAP partition of $x_1, \ldots, x_n$ in the conjugate exponential Bayesian Mixture Model, where the hyperparameter is identifiable. Then the family $\{\boldsymbol{x}_I \colon I \in \hat{\mathcal{I}}\}$ is T-linearly separable.*

This separation result for the clusters of the MAP partition implies, loosely speaking, that the MAP clusters are contained within some decent 'chunks' of the observation space. This motivates us to 'reverse the optics' and consider clusterings (that we call *induced*) of the data that are defined by an a'priori fixed partition $\mathcal{A}$ of the observation space. We derive the asymptotic limit of the logarithm of the posterior probability (up to a norming constant) of such induced clusterings, when the data are sampled independently from some given probability P (we call it *the input probability*). The result clearly depends on $\mathcal{A}$ and P. The limit is denoted by $\Delta_P^{\mathcal{M}}(\mathcal{A})$, where $\mathcal{M}$ represents the conjugate exponential family used to build the model. The limit does not depend on the exact specification of the prior distribution π on the component probabilities (cf. (1)), provided that π has a full support on the infinitely dimensional simplex $\triangle^\infty$.

3.2 Induced partitions and the limit formula

In this section we assume that the data is an independent sample from some fixed probability distribution P on $\mathbb{R}^d$, which we will call *the input distribution.* With the partition of the observation space fixed, this gives a random sequence of the clustering of indices, which in turn can be scored by the 'posterior score' (5). In the following we derive the asymptotic behaviour of the score. Note that, in the derivation of the model, the observations are not produced by an (unconditionally) i.i.d. sampling. This (of course) does not imply any 'mis-specification' if we derive asymptotic formulae by considering $X_1, X_2, \ldots$ as i.i.d. P random vectors; if P_n is the empirical distribution where n observations are generated using the scheme of the previous section, then $P_n \xrightarrow[(d)]{n\to\infty} P$ for some P and, for asymptotic results, the Strong Law of Large Numbers gives that the same asymptotics will hold for $X_1, X_2, \ldots$ i.i.d. P.

Of course, only a small class of distributions P can be generated according to the sampling scheme; these will necessarily be infinite mixtures of exponential distributions (and the mixture will have an *infinite* number of components). We do not limit ourselves to P that can be generated in this way and we consider more general input distributions in our analysis of the performance of the classifier.

Definition 3.4. Let P be a probability distribution on $\mathbb{R}^d$. We say that a family $\mathcal{A}$ of P-measurable subsets of $\mathbb{R}^d$ is a *P-partition* if

- $P(A) > 0$ for all $A \in \mathcal{A}$,
- $P\left(\bigcup_{A\in\mathcal{A}} A\right) = 1$,
- $P(A \cap B) = 0$ for all $A, B \in \mathcal{A}$, $A \neq B$.

Notation. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a sequence of vectors in $\mathbb{R}^d$. Let $\mathcal{A}$ be a countable collection of subsets of $\mathbb{R}^d$. We denote $\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{x}) := \{J_n^A \colon A \in \mathcal{A}\}$ where $J_n^A = \{i \leq n \colon X_i \in A\}$ (if $J_n^A = \emptyset$, we do not include it in $\mathcal{I}_n^{\mathcal{A}}$). If every x_i belongs to exactly one $A \in \mathcal{A}$ then $\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{x})$ is a partition of $[n]$. We say that it is *induced by* $\mathcal{A}$. The argument $\boldsymbol{x}$ is often clear from the context and therefore it is sometimes omitted.

Remark 3.5. It is clear by the definition of the P partition that if $\mathcal{A}$ is P-partition and $X_1, X_2 \ldots \overset{\text{iid}}{\sim} P$ then almost surely $\mathcal{I}_n^{\mathcal{A}}(X_1, \ldots, X_n)$ is a partition of $[n]$ for every $n \in \mathbb{N}$.

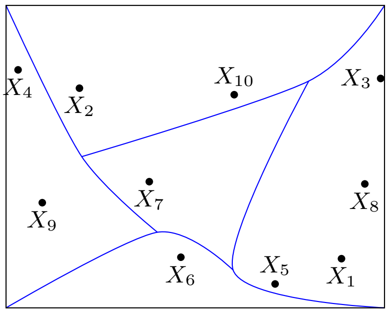


Figure 2: In this picture the observation space $\mathcal{X}$ is the rectangle and the partition $\mathcal{A}$ is defined by the blue separation curves. The points $X_1, \dots, X_{10}$ are drawn uniformly from $\mathcal{X}$. The random partition of $\{1, 2, \dots, 10\}$ induced by $\mathcal{A}$ is

$$\{\{1,3,5,8\},\{2,10\},\{4,9\},\{6\},\{7\}\}.$$

According to Remark 3.5, partitions induced by a P-partition on a random sample from P are almost surely partitions, and hence we can analyse their posterior probability in the conjugate exponential Bayesian Mixture models. We investigate the asymptotic limit of the logarithm of the joint probability given by (5). In order to specify the limit, we recall the notion of *convex conjugate.*

Definition 3.6. If f is a real function on $\mathbb{R}^d$ then the *convex conjugate* of f is the function $f^*\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, given by $f^*(\boldsymbol{z}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \big(\boldsymbol{z}^\top \boldsymbol{x} - f(\boldsymbol{x})\big)$.

Theorem 3.7. *Consider the infinite conjugate exponential Bayesian Mixture Model, in which the component measures are given by* (7) *and the base measure is given by* (11)*. Suppose that the exponential family is regular and that the convexity assumption* (13) *holds. Let P be a probability distribution on $\mathbb{R}^d$, $\mathcal{A}$ be a finite P-partition of $\mathbb{R}^d$ and $X \sim P$. Assume that $E_P \log h(X) < \infty$, $E_P\|T(X)\| < \infty$ and*

(i) $\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) < \infty$, *where A^* is the convex conjugate of the log-partition function A, given by* (9)*,*

(ii) $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big) \in \operatorname{int}\Omega$ *for some $r \in \mathbb{N}$, where Ω is the natural hyperparameter space.*

Let Q be the joint probability function given by (5)*, in which $g_{\vartheta,k}$ is given by* (12)*. Let $X_1, X_2, \ldots \overset{iid}{\sim} P$ and $\mathbf{X}_{1:n} = (X_1, \dots, X_n)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}}(\mathbf{X}_{1:n})\big) \overset{a.s.}{=} E_P \log h(X) + \Delta_P(\mathcal{A}) \tag{16}$$

where

$$\Delta_P(\mathcal{A}) = \sum_{A\in\mathcal{A}} P(A)\cdot \mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) + \sum_{A\in\mathcal{A}} P(A) \log P(A). \tag{17}$$

From this it can be deduced that the asymptotic formula for the Normal-Normal model has the form

$$\Delta_P^{NN}(\mathcal{A}) = \frac{1}{2} \sum_{A\in\mathcal{A}} P(A) \cdot \|E_P(RX \,|\, X \in A)\|^2 + \mathcal{H}(\mathcal{A}). \tag{18}$$

and the asymptotic formula for the Normal-Inverse-Wishart is

$$\Delta_P^{NIW}(\mathcal{A}) = -\frac{d}{2} - \frac{1}{2}\sum_{A\in\mathcal{A}} P(A)\cdot\log|\mathbf{V}_P(X\mid X\in A)| + \mathcal{H}(\mathcal{A}). \tag{19}$$

Let us point out an obvious consequence of Theorem 3.7.

Corollary 3.8. *Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two finite P-partitions of $\mathbb{R}^d$ such that $\Delta_P(\mathcal{A}_1) > \Delta_P(\mathcal{A}_2)$. Let $X_1, X_2, \ldots \overset{iid}{\sim} P$ and let $\mathbf{X}_{1:n} = (X_1,\ldots,X_n)$. With the assumptions of Theorem 3.7 almost surely there exists N such that*

$$Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}_1}(\mathbf{X}_{1:n})\big) > Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}_1}(\mathbf{X}_{1:n})\big) \quad \textit{for } n > N \tag{20}$$

Hence, as long as the induced partitions are concerned, the Δ_P function is an indicator of which of these partitions gives larger posterior score given by (5), when our data is an independent sample from the probability distribution P. In this sense we can hope that Δ_P relates somehow to the search of the MAP clustering. Clearly, the MAP clustering is not an induced one, but since the clusters in this case can be separated by some regular surfaces (cf. Theorem 3.3), we can hope that in the limit the MAP clustering can manifest some 'induced' behaviour. This idea is successfully applied in Rajkowski (2019) in a very specific setting of Normal-Normal model and the Chinese Restaurant prior on the space of partitions. This is described in more detail in the folowing subsection.

3.3 Asymptotic results for the Normal-Normal model

Consider the Normal-Normal BMM and let P be some input distribution with a bounded support, continuous with respect to the Lebesgue measure. Let $X_1, X_2 \ldots \sim P$, $\hat{\mathcal{I}}_n$ be the MAP partition of $X_1,\ldots,X_n$ and let $\hat{\mathcal{A}}_n = \big\{\operatorname{conv}\{\mathbf{X}_j : j\in I\} : I \in \hat{\mathcal{I}}_n\big\}$, where $\operatorname{conv} A$ is the convex hull of the set A. Let $\boldsymbol{M}_\Delta$ be the set of all P-partitions that maximise the Δ_P^{NN} function. Let d_P be the symmetric difference metric (i.e. for two P-measurable sets A, B we have $d_P(A,B) = (A\setminus B)\cup(B\setminus A)$) and let $\overline{d_P}$ be its natural extension to finite P-partitions.

The aforementioned limit result can be expressed in the following

Proposition 3.9. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then $\boldsymbol{M}_\Delta \neq \emptyset$ and almost surely $\inf_{\mathcal{M}\in\boldsymbol{M}_\Delta} \overline{d_P}(\hat{\mathcal{A}}_n, \mathcal{M}) \to 0$.*

It can be shown that as the norm of the within group covariance matrix tends to 0, the variance of the conditional expected value gains larger importance in maximising the function Δ_P^{NN} in formula (18) and this variance increases as the number of clusters increases. Therefore by manipulating the within group covariance parameter, when the input distribution is bounded it is possible to obtain an arbitrarily large (but fixed) number of clusters in the MAP partition as $n\to\infty$, as Theorem 3.10 states. This is also an indication of the inconsistency of the procedure used since it implies that when the input comes from a finite mixture of distributions with bounded support, then setting the Σ parameter too small leads to an overestimation of the number of clusters. This corresponds to some extent to the starting point of our research, which was the inconsistency result for the number of clusters of Miller and Harrison (2014).

Theorem 3.10. *Assume that P has bounded support and is continuous with respect to Lebesgue measure and let $X_1, X_2, \ldots \overset{iid}{\sim} P$. Then almost surely for every $K \in \mathbb{N}$ there exists an $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ such that if $\|\Sigma_0\| < \varepsilon$ and $n > n_0$ then $|\hat{\mathcal{I}}_n(\mathbf{X}_{1:n})| > K$.*

4 The adjusted Normal-Inverse-Wishart model

In the dissertation we consider the 'uniform input distribution' case and establish what partitions of the $[0,1]$ segments maximise the Δ_P function for the Normal-Normal and Normal-Inverse-Wishart model (given by (18)) and (19)) when the input distribution P is uniform on $[0,1]$. In the Normal-Normal case the within-cluster covariance is strongly influenced by the prior covariance parameter; the maximiser is unique and it is a division into segments of equal length, that make the within cluster covariance as close as possible to the value of the parameter (cf. Proposition 2.30 in the dissertation). When in the 'real' clustering the covariance is not the same for each cluster, or if the 'correct' hyper parameter value is not known in advance, then this model performs poorly; Proposition 3.24 illustrates that under hyperparameter misspecification, the model can behave very poorly.

To circumvent this, we place an Inverse Wishart prior over the within-cluster covariance parameter, but the naive application of such a prior produces a model which, when applied to a uniform input distribution, gives the same maximising value for the objective for any division of $[0,1]$ into connected pieces. The problem is that the parameter space for this non-parametric Bayes model is too large. Hence, we investigate priors which have a regularising effect; to obtain a suitable objective as an asymptotic limit, we consider prior distributions which depend on the number of observations.

It turns out that the only dependence on n which gives the regularising effect that we require is the Normal-Inverse-Wishart model with $\nu_0 = \alpha + \lambda n$ for parameters α and λ, while keeping the *expected* within cluster covariance fixed as Σ_0. More explicitly, we consider the asymptotic limit when, for a sample size n, the prior is

$$\Lambda \sim W^{-1}\big(\alpha + \lambda n + d + 1, (\alpha + \lambda n)\Sigma_0\big) \qquad \mu \,|\, \Lambda \sim N\big(\mu_0, \frac{1}{\kappa_0}\Lambda\big). \tag{21}$$

This leads to a parametrised family of objectives, which depend on the parameter λ. For fixed Σ_0 letting λ range between 0 and $+\infty$ gives a whole range of objectives, where $\lambda = +\infty$ corresponds to the situation of the previous chapter, where the within-cluster covariance is fixed as Σ_0. When $\lambda > 0$ (inequality strict), we can adapt the methods from Rajkowski (2019) (with fixed within-cluster covariance) and prove corresponding results.

By introducing a linear dependence of the concentration parameter in the Normal-Inverse-Wishart model on the number of observation we allowed to pass more prior knowledge about then within-cluster covariance structure into the model. In this setting we are able to show that in the MAP clustering for an infinite and bounded sequence of data, the size of clusters grows proportionally with the number of observations and, in turn, the number of clusters is bounded (Proposition 4.6 and Corollary 4.7 in the dissertation). We also compute the asymptotic limit of the posterior of an induced partition (an analogue of (19)) and we establish some properties of the limit (like monotonicity with respect to the λ parameter, cf. Lemma 4.4 in the dissertation). Finally, we suggest how to use the empirical of the limit to choose among various clustering proposals, which can be of interest to the practitioner. We finish the dissertation with the experimental analysis of this approach.

References

Jeffrey W. Miller and Matthew T. Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

Łukasz Rajkowski. Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Analysis*, 14(2):477–494, 2019.