Bayesian Analysis (2019) **14**, Number 2, pp. 477–494

Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model

Łukasz Rajkowski*

Abstract. Mixture models are a natural choice in many applications, but it can be difficult to place an a priori upper bound on the number of components. To circumvent this, investigators are turning increasingly to Dirichlet process mixture models (DPMMs). It is therefore important to develop an understanding of the strengths and weaknesses of this approach. This work considers the MAP (maximum a posteriori) clustering for the Gaussian DPMM (where the cluster means have Gaussian distribution and, for each cluster, the observations within the cluster have Gaussian distribution). Some desirable properties of the MAP partition are proved: 'almost disjointness' of the convex hulls of clusters (they may have at most one point in common) and (with natural assumptions) the comparability of sizes of those clusters that intersect any fixed ball with the number of observations (as the latter goes to infinity). Consequently, the number of such clusters remains bounded. Furthermore, if the data arises from independent identically distributed sampling from a given distribution with bounded support then the asymptotic MAP partition of the observation space maximises a function which has a straightforward expression, which depends only on the within-group covariance parameter. As the operator norm of this covariance parameter decreases, the number of clusters in the MAP partition becomes arbitrarily large, which may lead to the overestimation of the number of mixture components.

MSC 2010 subject classifications: 62F15.

Keywords: Dirichlet process mixture models, Chinese Restaurant Process.

1 Introduction

1.1 Motivation and new contributions

Clustering is a central task in statistical data analysis. A Bayesian approach is to model data as coming from a random mixture of distributions and derive the posterior distribution on the space of possible divisions into clusters. When there is not a natural a priori upper bound on the number of clusters, an increasingly popular technique to use is Dirichlet Process Mixture Models (DPMMs). It is therefore important to develop an understanding of the strengths and weaknesses of this approach.

Miller and Harrison (2014) restrict attention to the number of clusters produced by such a procedure and are somewhat critical of the method. Their main result implies that in a very general setting, the Bayesian posterior estimate of the *number* of clusters

*University of Warsaw, Faculty of Mathematics, Informatics and Mechanics. Banacha 2, 02-097 Warsaw, Poland, l.rajkowski@mimuw.edu.pl

 https://doi.org/10.1214/18-BA1114

is not consistent, in the sense that for any $t \in \{1, 2, \ldots\}$ almost surely

$$\limsup_{n\to\infty} \mathbb{P}(T_n = t | X_1, \ldots, X_n) < 1,$$

where $X_1, X_2, \ldots$ is an i.i.d. sample from a mixture with t components and T_n denotes the number of clusters to which the data are assigned. Here $\mathbb{P}$ is the probability in the probability space on which $X_1, X_2, \ldots$ are defined.

The Miller and Harrison inconsistency result relates to estimation of the *number* of clusters, not the classification itself. While they do not pursue this, they do provide examples of the structure estimators, such as the *MAP* (maximal a posteriori) partition, which maximises the posterior probability and the *mean partition*, introduced in Huelsenbeck and Andolfatto (2007), which minimises the sum of the squared distance between the mean partition and all partitions sampled by the MCMC (Markov Chain Monte Carlo) algorithm which they run, where the *distance* is the minimum number of individuals that have to be deleted from both partitions to make them the same.

This article presents developments that concern the properties of the MAP estimator in a *Gaussian* mixture model, where the cluster *centres* are generated according to a Gaussian distribution and, conditioned on the cluster centre, the observations within a cluster are generated by Gaussian distribution. The clusters are generated according to a Dirichlet Process. Analysing the MAP partition seems to be a natural choice. It is listed, for example, in Fritsch et al. (2009) as an established method. Of course, the set of possible candidates for the maximiser has to be restricted, since the space of partitions is too large for an exhaustive search. For example, Dahl (2006) suggests choosing the MAP estimator from a sample from the posterior. He notes, however, a potential problem of this approach; there may be only a small difference in the posterior probability between two significantly different partitions. This may indicate that the classifier is giving the wrong answer as a consequence of mis-specification of the within-cluster covariance parameter. We investigate such instability in our examples.

The conclusions of our analysis may be summarised as follows:

1. The convex hulls of the clusters are pairwise 'almost disjoint' (they may have at most one point in common, which must be a data point).
2. The clusters are of reasonable size; if $(\frac{1}{n}\sum_{j=1}^{n} \|x_j\|^2)_{n=1}^{\infty}$ (the sequence of means of squared Euclidean norms) is bounded, then $\liminf_{n\to\infty} \frac{m_n^{[r]}}{n} > 0$ for any $r > 0$, where $m_n^{[r]}$ denotes the number of observations in the smallest cluster (in the MAP partition of the first n observations) with non-empty intersection with $B(\mathbf{0}, r)$ (the ball of radius r, centred at the origin).
3. This implies that for any $r > 0$ the number of clusters in the n-th MAP partition required to cover observations inside $B(\mathbf{0}, r)$ remains bounded as $n \to \infty$.
4. When the data sequence comes from an i.i.d. sampling with bounded support there is an elegant formula to describe the limit of the MAP clustering; it is the partition of the observation space that maximises the function Δ given by (6).

In general, though it is a hard problem to find the global maximiser for this expression. Furthermore, the only parameter that this function depends on is the within-group covariance parameter.

5. The *negative* finding of the paper is that the clustering is very sensitive to the specification of the within-cluster variance and model mis-specification can lead to very misleading clustering. For example, if the data is i.i.d. from an input distribution which is uniform over a ball of radius r in $\mathbb{R}^2$ and the within-cluster variance parameter is $\sigma^2 I$, then for small σ, the classifier partitions the ball into several, seemingly arbitrary, convex sets. This classifier therefore has to be treated with caution.

1.2 Organisation of the article

We now present a brief overview of the structure of the paper. In Section 2 we give key definitions and provide complete mathematical statements of the main results together with intuitive explanations. Section 3 presents examples which illustrate the results obtained in the article. These examples show the MAP clustering obtained in various situations where the data comes from i.i.d. sampling. They indicate that this procedure may fail to produce reasonable output. The examples are supported by numerical simulations, which are described in Supplement B (Rajkowski, 2018). Section 4 contains a detailed presentation of the asymptotic proposition together with some related developments. In Section 5 we state the open problems and plans for future work.

2 Main results

2.1 The model

This section presents definitions of fundamental notions of our considerations together with some of their basic properties and relevant formulas. We show how they can be used to construct a statistical model in which we expect the data to be generated from different sources of randomness, without an a priori upper bound on the number of these sources a priori. We start with the definition of the Dirichlet Process, formally introduced in Ferguson (1973).

Definition. Let Ω be a space and $\mathcal{F}$ a σ-field of its subsets. Let $\alpha > 0$ and G_0 be a probability measure on $(\Omega, \mathcal{F})$. The *Dirichlet Process* on Ω with parameters α and G_0 is a stochastic process $(G(A))_{A\in\mathcal{F}}$ such that for every finite partition $\{A_1, \ldots, A_p\} \subseteq \mathcal{F}$ of Ω the random vector $(G(A_1), \ldots, G(A_p))$ has Dirichlet distribution with parameters $\alpha G_0(A_1), \ldots, \alpha G_0(A_p)$. In this case we write $G \sim \mathrm{DP}(\alpha, G_0)$.

As considered in Antoniak (1974), the Dirichlet Process can be used to construct a mixture model in which the number of clusters is not known a priori. The details are given in the following definition.

Definition. Let $(\Theta, \mathcal{F})$ be the parameter space and $(\mathcal{X}, \mathcal{B})$ the observation space. Let $\alpha > 0$ and G_0 be a probability measure on $(\mathcal{X}, \mathcal{F})$. Let $\{F_\theta\}_{\theta\in\Theta}$ be a family of probability

distributions on $(\mathcal{X}, \mathcal{B})$. The *Dirichlet Process mixture model* is defined by the following scheme for generating a random sample from the space $(\mathcal{X}, \mathcal{F})$

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \,|\, G &\overset{\text{iid}}{\sim} G \\ x_i \,|\, \boldsymbol{\theta}, G &\sim F_{\theta_i} \qquad \text{independently for } i \leq n. \end{aligned} \tag{1}$$

In Blackwell and MacQueen (1973) it is shown that the first two stages of (1) may be replaced by the following recursive procedure:

$$\theta_1 \sim G_0, \quad \theta_i \,|\, \theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{j=1}^{i-1} \frac{1}{\alpha + i - 1} \delta_{\theta_j}, \tag{2}$$

where δ_θ is the probability measure that assigns probability 1 to the singleton $\{\theta\}$. Of course, this procedure can be used to generate sequences of arbitrary length; the distribution of the resulting infinite sequence $(\theta_i)_{i=1}^{\infty}$ produced in this way is called the *Hoppe urn scheme*. Note that a realisation of this scheme defines a partition of $\mathbb{N}$ by a natural equivalence relation $(i \sim j) \equiv (\theta_i = \theta_j)$. Restriction of this partition to sets $[n]$ for $n \in \mathbb{N}$ is called the *Chinese Restaurant Process* (the CRP, for short).

Definition. The *Chinese Restaurant Process* with parameter α is a sequence of random partitions $(\mathcal{J}_n)_{n \in \mathbb{N}}$, where $\mathcal{J}_n$ is a partition of $[n] = \{1, 2, \ldots, n\}$, that satisfies

$$\mathcal{J}_{n+1} \,|\, \mathcal{J}_n = \{J_1, \ldots, J_k\} \sim \begin{cases} \{J_1, \ldots, J_i \cup \{n+1\}, \ldots, J_k\} & \text{with probability } \frac{|J_i|}{n+\alpha} \\ \{J_1, \ldots, J_k, \{n+1\}\} & \text{with probability } \frac{\alpha}{n+\alpha} \end{cases}. \tag{3}$$

We write $\mathcal{J}_n \sim \mathrm{CRP}(\alpha)_n$.

The Dirichlet Process mixture model for n observations is therefore equivalent to

$$\begin{aligned} \mathcal{J} &\sim \mathrm{CRP}(\alpha)_n \\ \boldsymbol{\theta} = (\theta_J)_{J \in \mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} G_0 \\ \mathbf{x}_J = (x_j)_{j \in J} \,|\, \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} F_\theta \qquad \text{for } J \in \mathcal{J}. \end{aligned} \tag{4}$$

We will refer to this formulation as the *CRP-based model*. In this paper we focus our attention on the Gaussian case, in which $\Theta = \mathbb{R}^d$, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F}$ and $\mathcal{B}$ are σ-fields of Borel sets, $G_0 = \mathcal{N}(\mu, \mathrm{T})$ and $F_\theta = \mathcal{N}(\theta, \Sigma)$ for $\theta \in \Theta$, where $\mu \in \mathbb{R}^d$ and $\mathrm{T}, \Sigma \in \mathbb{R}^{d,d}$ are the parameters of the model. This will be called the *CRP-based Gaussian model*. We also limit ourselves to the case where $\mu = 0$, however it may be easily seen that this is not a real restriction; the sampling from the zero-mean Gaussian model and transposing the output by the vector μ is equivalent to sampling from the Gaussian model with mean μ. Therefore all the clustering properties of the model can be investigated with the assumption that $\mu = 0$.

Remark 1. *The conditional probability of partition $\mathcal{J}$ in the zero-mean Gaussian model, given the observation vector $\mathbf{x} = (x_j)_{j=1}^n$, is proportional to*

$$C^{|\mathcal{J}|} \prod_{J\in\mathcal{J}} \frac{|J|!}{|J|^{(d+2)/2} \det R_{|J|}} \cdot \exp\left\{\frac{1}{2}\sum_{J\in\mathcal{J}} |J| \cdot \left\|R_{|J|}^{-1} R^2 \overline{\mathbf{x}_J}\right\|^2\right\} =: Q_{\mathbf{x}}(\mathcal{J}), \tag{5}$$

where $C = \alpha/\sqrt{\det T}$, $R = \Sigma^{-1/2}$, $R_m = (\Sigma^{-1} + T^{-1}/m)^{1/2}$ *for* $m \in \mathbb{N}$, $\|\cdot\|$ *is the standard Euclidean norm in* $\mathbb{R}^d$ *and* $\overline{\mathbf{x}_J} = \frac{1}{|J|}\sum_{j\in J} x_j$.

Proof. See Supplement A (Rajkowski, 2018). □

Having established the model we are now able to use it for inference about the data structure. A natural choice is to choose the partition that maximises the posterior probability given by (5). This leads to the notion of the MAP partition.

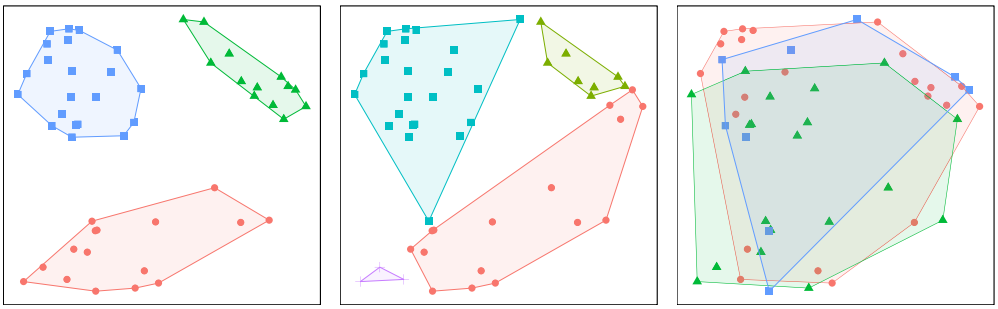
Definition. The *maximal a posteriori* (MAP) partition of $[n]$ with observed $\mathbf{x} = (x_i)_{i=1}^n$ is any partition of $[n]$ that maximises $Q_{\mathbf{x}}(\cdot)$ (equivalently, the posterior probability). We denote a maximiser by $\hat{\mathcal{J}}(\mathbf{x})$ (note: a priori this may not be unique).

2.2 Results

The first result is Proposition 1 which states that the MAP partition divides the data into clusters whose convex hulls are disjoint, with the possible exception of one datum.

Proposition 1. *For every* $n \in \mathbb{N}$ *if* $J_1, J_2 \in \hat{\mathcal{J}}(x_1, \ldots, x_n)$, $J_1 \neq J_2$ *and* A_k *is the convex hull of the set* $\{x_i : i \in J_k\}$ *for* $k = 1, 2$ *then* $A_1 \cap A_2$ *is an empty set or a singleton* $\{x_i\}$ *for some* $i \leq n$.

Proof. See Supplement A. □



(a) This is the desired partition which is also convex.

(b) This is a convex partition which is not ideal.

(c) This partition is not convex and it is clearly a bad one.

Figure 1: Illustration of the convexity property of a partition of the data. Clusters are indicated by the shape and colour of the points.

We say that a partition satisfying the property described by Proposition 1 is a *convex partition*. As Figure 1 indicates, this is a rather desirable feature of a clustering mechanism.

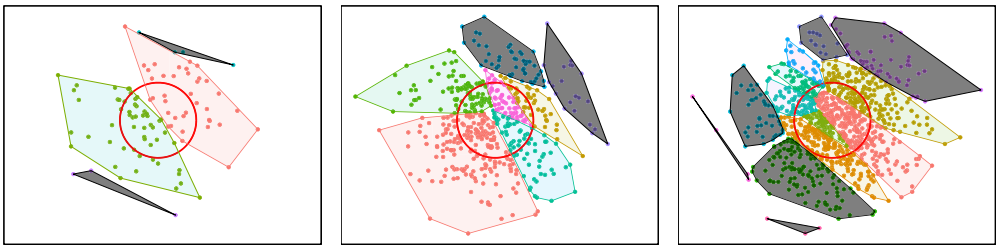
The next development give information about the size and number of the clusters. Proposition 2 states that when the sequence of sample 'second moments' is bounded then the size of the smallest cluster in the MAP partition among those that intersect a ball of given radius is comparable with the sample size.

Proposition 2. *If* $\sup_n \frac{1}{n}\sum_{i=1}^n \|x_n\|^2 < \infty$ *then*

$$\liminf_{n\to\infty} \min\{|J|\colon J \in \hat{\mathcal{J}}(x_1,\ldots,x_n), \exists_{j\in J}\|x_j\| < r\}/n > 0$$

for every $r > 0$.

Proof. See Supplement A. □



(a) $n = 100$ (b) $n = 500$ (c) $n = 1000$

Figure 2: Illustration of Proposition 2 and Corollary 2. The red circle is arbitrarily fixed and the clusters it intersects are coloured. The number of observations in each coloured cluster is proportional to n and the number of these clusters remains bounded as $n \to \infty$.

The assumption $\sup_n \frac{1}{n}\sum_{i=1}^n \|x_n\|^2 < \infty$ allows the data sequence to be unbounded but it does ensure that it does not grow too quickly. It is easy to see that an assumption of this kind is necessary, otherwise it would be possible for each new observation to be large enough to create a new singleton cluster.

A simple consequence of Proposition 2 is that under these assumptions the number of components in the MAP partition that intersect a given ball is almost surely bounded.

Corollary 2. *If* $(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2)_{n=1}^\infty$ *is bounded then for every* $r > 0$ *the number of clusters that intersect* $B(\mathbf{0}, r)$ *is bounded, i.e.*

$$\limsup_{n\to\infty} |\{J \in \hat{\mathcal{J}}(x_1,\ldots,x_n)\colon \exists_{j\in J}\|x_j\| < r\}| < \infty.$$

Proof. The proof follows easily from the fact that the size of the smallest cluster that intersects $B(\mathbf{0}, r)$ is bounded from above by the number of observations divided by the number of clusters intersecting the ball. □

In order to formulate the central result of the paper we need to introduce several notions. Let P be a probability distribution on $\mathbb{R}^d$ and X a random variable with distribution P. Let Δ be the function on the space of finite families of measurable sets defined by the following formula

$$\Delta(\mathcal{G}) = \frac{1}{2}\sum_{G\in\mathcal{G}} P(G)\left\|R\mathbb{E}\left(X \mid X \in G\right)\right\|^2 + \sum_{G\in\mathcal{G}} P(G)\ln P(G), \tag{6}$$

where R^2 is the inverse of the within-cluster covariance matrix Σ and $\mathbb{E}\left(X \mid X \in G\right)$ is the expected value of X conditioned on $X \in G$.

We consider the *symmetric distance metric* over P-measurable sets, which is defined by $d_P(A,B) = P((A \setminus B) \cup (B \setminus A))$. This can be easily extended to a metric $\overline{d_P}$ over finite families of measurable subsets of $\mathbb{R}^d$ (details are given in Section 4.3). Also we say that a family of measurable sets $\mathcal{A}$ is a *P-partition* if $P(\bigcup_{A\in\mathcal{A}} A) = 1$ and $P(A\cap B) = 0$ for all $A, B \in \mathcal{A}$, $A \neq B$. Let $\boldsymbol{M}_\Delta$ denote the set of finite P-partitions that maximise the function Δ.

Consider $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ and let $\hat{\mathcal{A}}_n$ be the family of the convex hulls of clusters of observations in $\hat{\mathcal{J}}(X_1, \ldots, X_n)$.

Proposition 3. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then $\boldsymbol{M}_\Delta \neq \emptyset$ and almost surely* $\inf_{\mathcal{M}\in\boldsymbol{M}_\Delta} \overline{d_P}(\hat{\mathcal{A}}_n, \mathcal{M}) \to 0$.

Proof. The proof follows from Theorem 14. See Supplement A for details. □

The function Δ does not depend on the concentration parameter α or the between-groups covariance parameter. It therefore follows, somewhat surprisingly, that in the limit the shape of the MAP partition does not depend on these two parameters.

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 in formula (6) 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 Proposition 4 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.

Proposition 4. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then for every $K \in \mathbb{N}$ there exists an $\varepsilon > 0$ such that if $\|\Sigma\| < \varepsilon$ then $|\hat{\mathcal{J}}_n| > K$ for sufficiently large n.*

Proof. See Supplement A. □

It is worth pointing out that Proposition 1 and Proposition 2 hold also for *finite* Gaussian mixture models with Dirichlet prior on the probabilities of belonging to a

given cluster. Proposition 3 also remains true with $\boldsymbol{M}_\Delta$ replaced by $\boldsymbol{M}_\Delta^K$ – the set of P-partitions with at most K clusters that maximise the function Δ, where K is the number of clusters assumed by the model. The details are left for Supplement A.

3 Examples

This section presents some examples which illustrate the main propositions of the article. In Section 3.1 we compute the convex partition that maximises Δ when P is a uniform distribution on the interval $[-1, 1]$. Section 3.2 gives an example of a distribution with well-defined moments, for which the maximiser of Δ necessarily has infinitely many clusters, although for any $r < \infty$, the number of clusters that intersect a ball of radius r is finite. This example illustrates the content of Theorem 6, where it is shown that with appropriate choice of model parameters, if the input distribution is exponential then the number of clusters in the sequence of MAP partitions becomes arbitrarily large. Section 3.3 investigates Gaussian mixture models; the MAP partition does not properly identify the two clusters when the mixture distribution is bi-modal. Finally, in Section 3.4 we consider the uniform distribution on the unit disc in $\mathbb{R}^2$. The partition maximising the function Δ cannot be obtained by analytical methods, but it may be approximated. The results approximate the optimal partition of the unit disc and illustrate the convexity of Proposition 1. All examples are substantiated with computer simulations, presented in the main text or in Supplement B.

3.1 Uniform distribution on an interval

We find the convex partition that maximises Δ if P is a uniform distribution on $[-1, 1]$. Firstly we find an optimal partition with fixed number of clusters K. Since it is convex, it is defined by the lengths of K consecutive subintervals of $[-1, 1]$. Let those be $2p_1, \ldots, 2p_n$. Computations in Supplement A show that with K fixed the optimal division is $p_1 = p_2 = \ldots = p_K = 1/K$. Using this, it is computed that the optimal number of clusters is $K = \lfloor R/\sqrt{3} \rfloor$ or $K = \lceil R/\sqrt{3} \rceil$, where $\lfloor x \rfloor$ and $\lceil x \rceil$ are the largest integer not greater than x and the smallest integer not less than x, respectively. It is worth noting that the variance of the data within a segment of length $2R/\sqrt{3}$ is equal to R, so in this case the MAP clustering splits the data in a way that adjusts the empirical within-group covariance to the model assumptions.

It should be underlined that in this example, if Σ is small, the MAP partition has more than one cluster. The clustering is therefore misleading, since in this case there is exactly one population (which is uniform $[-1, 1]$). The number of clusters in the MAP partition becomes arbitrarily large as Σ goes to 0, as Proposition 4 states.

This would suggest that, in general, a sensible choice of Σ should be made a priori. The sample variance would give an upper bound on Σ (since the data variance is the sum of between-group and within-group variances), but there is no natural lower bound for this parameter. In this example the partitioning mechanism itself is clearly far from satisfactory when it produces more than two clusters; the divisions seem very arbitrary.

3.2 Exponential distribution

When the input distribution is exponential with parameter 1, then for a relevant choice of model parameters (e.g. $\alpha = \mathrm{T} = 1$, $\Sigma = 4$) there is no finite partition that maximises Δ; the value of the function Δ for a given convex partition can be increased by taking any interval of length larger than 3 and dividing it into two equally probable parts. See Supplement A for the proof.

Since the exponential distribution does not have bounded support, our considerations regarding the relation between the function Δ and the MAP clustering cannot be applied directly. However, by using similar methods we can establish that for exponential input the MAP procedure creates an arbitrarily large number of clusters. This is stated in Theorem 6, whose proof is presented in Supplement A.

3.3 Mixture of two normals

Let the input distribution be a mixture of two normals ($P = \frac{1}{2}(\nu_{-1.01} + \nu_{1.01})$), where ν_m is the normal distribution with mean m and variance 1). It can be proved that this distribution is bi-modal (however slightly; see Supplement A). Choose the model parameters consistent with the input distribution, i.e. $d = \alpha = \Sigma = \mathrm{T} = 1$. It can be computed numerically that $\Delta(\{(-\infty, 0], (0, \infty)\}) \approx -0.0046 < 0 = \Delta(\{\mathbb{R}\})$. An intuitive partition of the data into positive and negative is induced by the partition $\{(-\infty, 0], (0, \infty)\}$ and hence, by Corollary 8, for sufficiently large data input the posterior score for the two clusters partition is smaller than the posterior score for a single cluster. This may be taken as an indication of inconsistency of the MAP estimator in this setting.

3.4 Uniform distribution on a disc

This gives an example of non-uniqueness of the optimal partition, since the family of optimal partitions is clearly invariant under rotation around $(0, 0)$. Let P be uniform distribution on $B(\mathbf{0}, 1)$. It can be easily seen that $\Delta(B(\mathbf{0}, 1)) = 0$. Let R be the identity matrix and let B_1^+ (B_1^-) be a subset of $B(\mathbf{0}, 1)$ with non-negative (negative) first coordinate. Then $\Delta(\{B_r^+, B_r^-\}) = 2r^2/9 - \ln 2$. Therefore, for sufficiently large r, a partition of $B(\mathbf{0}, 1)$ into halves is better than a single cluster, hence the optimal convex partition $\mathcal{E}$ is not a single cluster. Since a single cluster is the only convex partition of $B(\mathbf{0}, 1)$ that is rotationally invariant about the origin, it follows that the optimal partition is not unique.

The simulation in this case also give a nice illustration of the convexity of the MAP partition, proved in Proposition 1 and show the arbitrary nature of the partitioning when r is large.

3.5 The MAP clustering properties

This short simulation study presents the performance of the MAP estimator when the input distribution is a mixture of uniform distributions on three pairwise disjoint

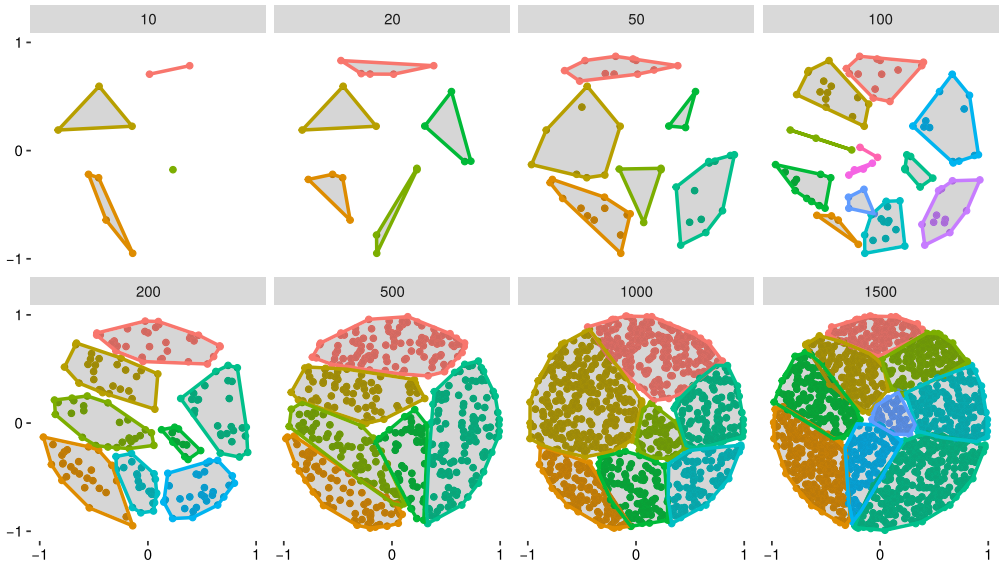


Figure 3: Clustering in the MAP partition of the first $k = 100, 500, 1000, 1500, 2000$ observations (in columns) in the i.i.d. sample from the uniform distribution of disc $B(\mathbf{0}, 1)$. Different clusters are denoted by different colours.

ellipses. The output is shown on Figure 4. It shows that the MAP clustering detects the mixture components or at least the clusters it creates are the sub-groups of the true mixture components (all depending on the within-group covariance parameter Σ). It also provides a nice illustration for two properties of the MAP partition: firstly the convexity property (Proposition 1) and secondly – the fact that when the within-group covariance parameter is decreasing, the number of cluster in the MAP partition grows, as stated in Proposition 4.

4 Detailed presentation of Proposition 3

4.1 Classification of randomly generated data

Let P be a probability distribution on $(\mathbb{R}^d, \mathcal{B})$ and let $(X_n)_{n=1}^{\infty}$ be a sequence of independent copies of a random variable X with distribution P. Then $\hat{\mathcal{J}}_n = \hat{\mathcal{J}}(X_1, \ldots, X_n)$ goes a random partition of $[n]$. Note that if $\mathbb{E}\,\|X\|^4 < \infty$ (here and subsequently, $\mathbb{E}$ denotes the expected value) then by the strong law of large numbers almost surely $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \to \mathbb{E}\,\|X\|^2 < \infty$ and therefore the assumptions of Proposition 2 are satisfied almost surely. Useful corollaries of this observation are listed below.

Corollary 3. *If* $\mathbb{E}\,\|X\|^4 < \infty$ *then for every* $r > 0$

(a) $\liminf_{n\to\infty} \min\{|J|\colon J \in \hat{\mathcal{J}}_n, \exists_{j\in J}\|X_j\| < r\}/n > 0$ *almost surely.*
(b) the number of clusters in $\hat{\mathcal{J}}_n$ *that intersect* $B(\mathbf{0}, r)$ *is bounded.*

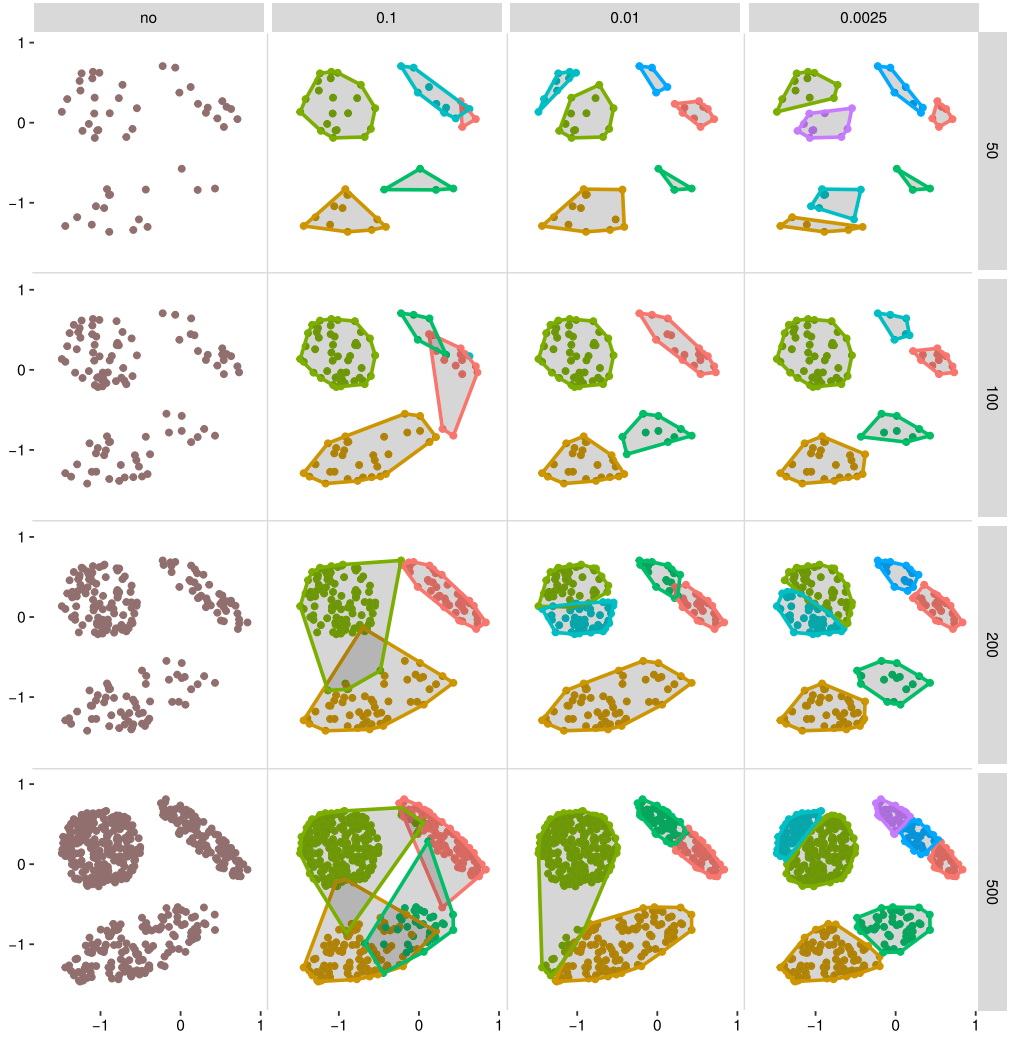


Figure 4: Clustering in the MAP partition of the first $k = 50, 100, 200, 500$ observations (in columns) in the i.i.d. sample from the mixture of three uniform distributions on a disjoint ellipses. The MAP clustering was constructed for $\alpha = 1$, $\mathrm{T} = I$ and $\Sigma = \sigma^2 I$ where $\sigma^2 \in \{1, .1, .01, .0025\}$ (in rows). Different clusters are denoted by different colours, the convex hulls of the clusters are also marked. It is clear that some of the partitions presented are not convex, particularly for large σ^2. This is due to the fact that the method is less than perfect. As σ^2 increases, the likelihood component of the formula for the posterior is less significant and hence partitions with the same prior (where clusters are of the same size) have similar posterior score. Therefore, with high probability, sampling from the posterior will not choose the MAP partition, or even a partition that reasonably resembles the MAP clustering. We mentioned this instability in Section 1.1.

An easy consequence of Corollary 3 is

Corollary 4. *If the support of P is bounded then*

(a) $\liminf_{n\to\infty} \min\{|J|\colon J \in \hat{\mathcal{J}}_n\}/n > 0$ *almost surely.*
(b) $|\hat{\mathcal{J}}_n|$ *is almost surely bounded.*

Proof. If the support of P is bounded then $\mathbb{E}\,\|X^4\| < \infty$. Therefore we can use Corollary 3 where we take r sufficiently large so that $B(\mathbf{0}, r)$ contains the support of P. □

The assumptions of Corollary 4 cannot be relaxed to those of Corollary 3. It turns out that there exists a probability distribution P with a countable number of atoms sufficiently far apart, whose probabilities are chosen so that $\mathbb{E}\,\|X\|^4 < \infty$ and almost surely the most recent observation creates a singleton in the sequence of MAP partitions infinitely often, i.e. there exists a sequence $(n_k)_{k=1}^{\infty}$ such that $\{x_{n_k}\} \in \hat{\mathcal{J}}_{n_k}$. This violates part (a) of Corollary 4. On the other hand, for appropriate parameter choice, sampling from the exponential distribution leads to the number of clusters in the MAP partition tending to infinity, which contradicts part (b) of Corollary 4. Proofs of these facts are left for Supplement A. These facts are now formally stated in the following two theorems:

Theorem 5. *If $d = 1$ and $\alpha = \mathrm{T} = \Sigma = 1$ then for $P = \sum_{m=0}^{\infty} q(1-q)^m \delta_{18^m}$, where $q = (2 \cdot 18)^{-1}$, almost surely $\liminf_{n\to\infty} m(\hat{\mathcal{J}}_n) = 1$.*

Theorem 6. *If $P = \mathrm{Exp}(1)$ and the CRP model parameters are $\alpha = \mathrm{T} = 1$, $\Sigma < (32 \ln 2)^{-1}$ then the number of clusters in the sequence of MAP partitions almost surely goes to infinity.*

4.2 The induced partition

Instead of searching for the MAP clustering, one may choose a simpler (and more arbitrary) way to partition the data. The idea is to choose a partition of the observation space in advance and then divide the sample assigning each datum to the element of this partition which contains it. We call this decision rule an *induced partition*. In this section we give a formal definition and investigate how it behaves when the input is identically distributed and how it relates to the formula for the posterior probability given by (5).

Definition. Let $\mathcal{A}$ be a fixed partition of $\mathbb{R}^d$. Let $J_n^A = \{i \le n\colon X_i \in A\}$ for $n \in \mathbb{N}$ and $A \in \mathcal{A}$ and define a random partition of $[n]$ by $\mathcal{J}_n^{\mathcal{A}} = \{J_n^A \neq \emptyset\colon A \in \mathcal{A}\}$. We say that this partition of $[n]$ is *induced by* $\mathcal{A}$.

In the following part of the text, for two sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ of nonzero real numbers, we use the notation $a_n \approx b_n$ to denote $\lim_{n\to\infty} a_n/b_n = 1$.

Lemma 7. *Let $\mathcal{A}$ be a finite P-partition of $\mathbb{R}^d$ consisting of Borel sets with positive P measure. Then almost surely $\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\mathcal{J}^{\mathcal{A}})} \approx \frac{n}{e}\exp\{\Delta(\mathcal{A})\}$, where Δ is the function defined by (6).*

Proof. See Supplement A. □

Corollary 8. *If $\mathcal{A}$, $\mathcal{B}$ are two finite P-partitions of $\mathbb{R}$ such that $\Delta(\mathcal{A}) > \Delta(\mathcal{B})$ then almost surely $Q_{\mathbf{X}_{1:n}}(\mathcal{J}_n^{\mathcal{A}}) > Q_{\mathbf{X}_{1:n}}(\mathcal{J}_n^{\mathcal{B}})$ for sufficiently large n.*

Proof. The proof is straightforward and therefore omitted. □

Corollary 8 implies that if we look for the optimal, finite induced partition, it will be a partition of the data induced by the finite partition of the observation space that maximises the function Δ. This formulation suggests a strong relationship between the MAP partition and the finite maximisers of Δ, which will be investigated further in Section 4.3, in the case where P has bounded support. The case where P does not have bounded support is beyond the scope of this work, for reasons presented in Section 5. This is a goal for future research.

At the end of this Section, let us provide an interpretation of the function Δ. Let $\mathcal{A}$ be a finite partition and $Z_{\mathcal{A}} = \mathbb{E}\,(X|\mathbf{1}_A(X)\colon A \in \mathcal{A})$ be the conditional expected value of X given the indicators $\mathbf{1}_A(X)$ for $A \in \mathcal{A}$. Then $Z_{\mathcal{A}}$ is a discrete random variable which is equal to $\mathbb{E}\,(X \,|\, X \in A)$ with probability $P(A)$. This implies that $\Delta(\mathcal{A}) = \frac{1}{2}\mathbb{E}\,\|RZ_{\mathcal{A}}\|^2 - H(Z_{\mathcal{A}})$, where the function H assigns to a random variable its entropy. Moreover

$$\mathbb{E}\,\|RZ_{\mathcal{A}}\|^2 = \operatorname{tr}\big(\mathbf{V}(RZ_{\mathcal{A}})\big) + \|\mathbb{E}\,RZ_{\mathcal{A}}\|^2 = \operatorname{tr}\big(R\mathbf{V}(Z_{\mathcal{A}})R^t\big) + \|R\mathbb{E}\,Z_{\mathcal{A}}\|^2$$

in which $\operatorname{tr}(\cdot)$ is the trace function and $\mathbf{V}(\cdot)$ is the covariance matrix of a given random vector. Since $\mathbb{E}\,Z_A = \mathbb{E}\,X$ we obtain that

$$\Delta(\mathcal{A}) = \frac{1}{2}\operatorname{tr}\big(R\mathbf{V}(Z_{\mathcal{A}})R^t\big) - H(Z_{\mathcal{A}}) + \frac{1}{2}\|R\mathbb{E}\,X\|^2. \tag{7}$$

Equation (7) justifies the following description of the function Δ: up to a constant, it may be treated as a difference between the variance and the entropy of the conditional expected value of a linearly transformed, P-distributed random variable given its affiliation to one of the sets in the partition.

4.3 Convergence of the MAP partitions

Corollary 8 gives us a convenient characterisation of the partitions of $\mathbb{R}^d$ that in the limit induce the best possible partitions of sets $[n]$. At this stage however we do not know yet if the best induced partitions relate to overall best partitions, namely the MAP partitions. A natural question is if the behaviour of the MAP partition resembles the induced classification introduced in Section 4.2, as the sample size goes to infinity, and under what conditions. This section presents partial answers in this regard; it should be stressed however that all the developments presented here are limited to the case when the input distribution has bounded support. The reasons for such limitation are briefly described in Section 5.

As we already know that clusters in the MAP partition create disjoint convex sets, the analysis of the approximate behaviour of these partitions would be easier if a form of 'uniform law of large numbers' with respect to the family of convex sets were true. More precisely if we let $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ we need the following to hold:

$$\lim_{n\to\infty} \sup_{C \text{ convex}} \left|P_n(C) - P(C)\right| = 0 \quad \text{almost surely.} \tag{$*$}$$

In other words we require that the class of convex sets is a *Glivenko–Cantelli class* with respect to P. A convenient condition for this to hold is given in Elker et al. (1979), Example 14:

Lemma 9. *If for each convex set C the boundary ∂C can be covered by countably many hyperplanes plus a set of P-measure zero, then ($*$) holds for P.*

In particular, it can easily be seen that the assumptions of Lemma 9 are satisfied if P has a density with respect to Lebesgue measure λ_d on $\mathbb{R}^d$ (since in this case the Lebesgue measure λ_d of the boundary of any convex set is 0, and hence is also P measure 0).

We can now formulate a functional relation between the posterior probability of the MAP partition and the value of the function Δ on the family of convex hulls of the sets in the MAP partition.

Lemma 10. *Assume that P has bounded support and satisfies ($*$). Let $\hat{\mathcal{A}}_n$ be the family of the convex hulls of the clusters in the MAP partition, i.e. $\hat{\mathcal{A}}_n = \{\mathrm{conv}\{\mathbf{X}_j \colon j \in J\} \colon J \in \hat{\mathcal{J}}\}$. Then almost surely*

$$\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\hat{\mathcal{J}}_n)} \approx \frac{n}{e}\exp\{\Delta(\hat{\mathcal{A}}_n)\}.$$

Proof. See Supplement A. □

Now we investigate the convergence of the sequence $\hat{\mathcal{A}}_n$ defined in Lemma 10. In order to do so we need a topology on relevant subspaces of $2^{\mathbb{R}}$. We begin by recalling two standard metrics used in this context.

Definition. Let $\mathcal{D}$ be a class of closed subsets of $\mathbb{R}^d$. Then the function $\varrho_H \colon \mathcal{D}^2 \to \mathbb{R}$ defined by

$$\varrho_H(A,B) = \inf\{\varepsilon > 0 \colon A \subseteq (B)_\varepsilon, B \subseteq (A)_\varepsilon\},$$

where $(X)_\varepsilon = \{x \in \mathbb{R}^d \colon \mathrm{dist}(x,X) < \varepsilon\}$, is a metric on $\mathcal{D}$. It is called the *Hausdorff distance*. The fact that it is a metric follows from 1.2.1 in Moszyńska (2005).

Definition. Let $\mathcal{M}$ be a σ-field on $\mathbb{R}^d$ and μ be a measure on $(\mathbb{R}^d, \mathcal{M})$. Then the function $d_\mu \colon \mathcal{M}^2 \to \mathbb{R}$ defined by $d_\mu(A,B) = \mu((A \setminus B) \cup (B \setminus A))$ is a pseudometric on $\mathcal{M}$, which by definition means that it is symmetric, nonnegative and satisfies the triangle inequality. It is called the *symmetric difference metric*. The fact that it is a pseudometric is explained in the beginning of Section 13, Chapter III of Doob (1994). Note that since $d_\mu(A,B) = 0$ does not imply $A = B$, formally d_μ is not a metric on $\mathcal{M}$. Although for our consideration the difference of measure 0 is of no importance, we keep on using the proper *pseudometric* term in this context.

The two following theorems are crucial for establishing the limits of maximisers. Theorem 11 is Theorem 3.2.14 in Moszyńska (2005); it ensures the existence of d_H-converging subsequence in every bounded sequence of convex sets. Theorem 12 is a straightforward consequence of Theorem 12.7 in Valentine (1964) (in the latter P is taken to be the Lebesgue measure). It states that when P has a density with respect to the Lebesgue measure then the Hausdorff metric restricted to $\mathcal{K}$ is stronger than the symmetric difference metric.

Theorem 11. *The space $(\mathcal{K}, \varrho_H)$ is finitely compact (i.e. every bounded sequence has a convergent subsequence).*

Theorem 12. *If P is continuous with respect to the Lebesgue measure then convergence in ϱ_H implies convergence in d_P in the space $\mathcal{K}$.*

Note that the Hausdorff and symmetric difference metrics are defined on sets. However we are interested in MAP partitions, which are *families* of sets. Therefore it is convenient to extend the definitions of these metrics to families of sets, as presented below. Remark 13 ensures that the desirable properties of compactness are preserved by such extension.

Definition. Let d be a pseudometric on the family of sets $\mathcal{F}$. For $K \in \mathbb{N}$ we define $F_K(\mathcal{F})$ to be the space of finite subfamilies of $\mathcal{F}$ that have at most K elements. Moreover $\mathcal{A} = \{A^{(1)}, \ldots, A^{(k)}\} \in F_K(\mathcal{F})$ and $\mathcal{B} = \{B^{(1)}, \ldots, B^{(l)}\} \in F_K(\mathcal{F})$ we define

$$\bar{d}(\mathcal{A}, \mathcal{B}) = \min_{\sigma \in \Sigma_K} \max_{i \leq K} d(A^{(i)}, B^{(\sigma(i))}), \tag{8}$$

where Σ_K is the set of all permutations of $[K]$ and we assume $A^{(i)} = \emptyset$ and $B^{(j)} = \emptyset$ for $i > k$ or $j > l$ respectively.

Remark 13. *If $(\mathcal{F}, d)$ is a pseudometric space then $(F_K(\mathcal{F}), \bar{d})$ is also a pseudometric space. Moreover, if $(\mathcal{F}, d)$ is finitely compact then $(F_K(\mathcal{F}), \bar{d})$ is also finitely compact.*

Proof. The proof is straightforward. See Supplement A for details. □

Now assume that P has bounded support. Then by Theorem 11 and Remark 13 it follows that $(\hat{\mathcal{A}}_n)_{n=1}^{\infty}$ has convergent subsequences which have a limit under $\overline{d_H}$ (note that as the support of P is bounded, sets $\hat{\mathcal{A}}$ are also bounded in the d_H metric). Let us denote the (random) set of their limits by $\boldsymbol{E}$. Note that by the properties of d_H distance each family in $\boldsymbol{E}$ consists of convex, closed sets. If we assume that P is continuous with respect to the Lebesgue measure then it follows from Lemma 10 together with Theorem 12 that $\boldsymbol{E}$ consists of finite P-partitions that maximise the function Δ.

Theorem 14. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then every partition in $\mathbf{E}$ is a finite P-partition that maximises Δ.*

Proof. See Supplement A. □

Now Proposition 3 is a straightforward, topological consequence of Theorem 14. This is shown in Supplement A.

5 Discussion

It should be clearly stated that the scope of the paper is limited in two ways. Firstly, only the Gaussian model is considered. It is natural to ask if the methods used here can be applied for other combinations of base measure and component distributions. The author is sceptical in this regard. The proofs of the key Proposition 1 and Proposition 2 rely strongly on the formula (5). It is difficult to find a computationally feasible choice of the base and component measures so that the resulting formula for the posterior probabilities has similar properties.

Secondly, the limiting results contained in Section 4.3 are proved in the case where the support of the input distribution is bounded. In this case the model is clearly misspecified. A significant effort was put in order to extend the results from Section 4.3 at least to the case where P is Gaussian. Unfortunately, there are some technical hurdles which the author was not able to overcome, which we now outline. The first result in which the boundedness of the input distribution is used is Lemma 10 – here we use both parts of Corollary 4 which, as shown by Theorem 5 and Theorem 6, cannot be easily generalised. A natural approach is to fix large $r > 0$ and use Corollary 3 – then the product of those factors in (5) which come from the clusters that intersect $B(\mathbf{0}, r)$ may be well approximated using Lemma 9, since by Corollary 3 there are finitely many clusters intersecting $B(\mathbf{0}, r)$ and the number of observations in the cluster is comparable with n for each cluster. Unfortunately in this way there is no control over the impact of the clusters outside $B(\mathbf{0}, r)$ as there are no lower bounds on their size and upper bounds on their number. However the author believes that these obstacles are possible to overcome and this remains subject for the future work.

It should be also underlined the setting of our analysis was not the usual one for the consistency analysis. Indeed, in our formulation of the CRP model our parameter space is the space of partitions of $[n]$, which is changing with n. To perform a classical consistency analysis we need the parameter space to be fixed regardless of the number of observations. On the other hand, if we consider the DPMM formulation, in which the parameter space is the space of all possible realisations of the Dirichlet Process (i.e. the space of discrete measures on $\mathbb{R}^d$ with infinitely many atoms) then again our input should come from an *infinite* mixture of normals, which was not the case in our examples.

However some of our results from Section 4 can be applied when the input sequence is a realisation of the DPMM. Indeed, the convexity result of Proposition 1 does not have any assumptions on the data sequence. As for Proposition 2, it requires the sequence of mean squared norms to be bounded. It is easy to prove (see Supplement A) that for a realisation of the DPMM this assumption holds almost surely and hence for every $r > 0$ the clusters intersecting $B(\mathbf{0}, r)$ in the sequence of the MAP partitions constructed on the sample from DPMM are of size comparable with the number of observation and their number is bounded. However, some fundamental questions remain unanswered in this case (e.g. does the number of clusters in the MAP partition tend to infinity in this case?) and they are open for further investigation.

Note that the machinery presented can be used for a different task. The P-partitions that maximise the function Δ seem to be interesting objects in their own right. Note

that for dimension greater than 1 it seems to be extremely difficult to derive the maximisers simply by analytical means. Remark 13 and Proposition 3 give us a convenient tool to examine those maximisers as they may be approximated by performing sampling from the posterior. This cannot be done faithfully as the normalizing constant in the formula (5) cannot be computed explicitly, however there are standard MCMC techniques that can be applied there (e.g. Neal (2000)). Further examination of the maximisers of the function Δ is left for future research.

Supplementary Material

Supplementary Material to "Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model" (DOI: 10.1214/18-BA1114SUPP; .zip). Supplement A: This supplementary material contains proofs that were left for the appendix. Supplement B: This supplementary material contains results of computer simulations.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 1152–1174. MR0365969. 479

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 353–355. MR0362614. 480

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." *Bayesian Inference for Gene Expression and Proteomics*, 201–218. MR2706330. 478

Doob, J. L. (1994). *Measure Theory*. Graduate Texts in Mathematics 143. Springer-Verlag New York, 1 edition. MR1253752. doi: https://doi.org/10.1007/978-1-4612-0877-8. 490

Elker, J., Pollard, D., and Stute, W. (1979). "Glivenko-Cantelli Theorems for Classes of Convex Sets." *Advances in Applied Probability*, 11(4): 820–833. MR0544197. doi: https://doi.org/10.2307/1426861. 490

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 209–230. MR0350949. 479

Fritsch, A., Ickstadt, K., et al. (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, 4(2): 367–391. MR2507368. doi: https://doi.org/10.1214/09-BA414. 478

Huelsenbeck, J. P. and Andolfatto, P. (2007). "Inference of population structure under a Dirichlet process model." *Genetics*, 175(4): 1787–1802. 478

Miller, J. W. and Harrison, M. T. (2014). "Inconsistency of Pitman-Yor Process Mixtures for the Number of Components." *Journal of Machine Learning Research*, 15: 3333–3370. MR3277163. 477

Moszyńska, M. (2005). *Selected Topics in Convex Geometry*. Birkhäuser Boston, 1 edition. MR2169492. 490, 491

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 493

Rajkowski, Ł. (2018). "Supplementary Material to "Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model"" *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1114SUPP. 479, 481

Valentine, F. A. (1964). *Convex Sets*. McGraw-Hill Book Company. MR0170264. 491

Acknowledgments

The author wishes to express his thanks to an anonymous Referee and the Associate Editor for their helpful comments and many thoughtful suggestions, which made the text far more readable and better than it was. Also, the author thanks dr John Noble for helpful discussions.