**University of Warsaw**
Faculty of Mathematics, Informatics and Mechanics

Łukasz Rajkowski

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

PhD dissertation

**Supervisors:**
prof. dr hab. Wojciech Niemiro
University of Warsaw
Nicolaus Copernicus University in Toruń

dr John Noble
University of Warsaw

March 2021

*Author's declaration*
I hereby declare that this dissertation is my own work.

2 March 2021 .....

Łukasz Rajkowski

*Supervisor's declaration*
This dissertation is ready to be reviewed.

2 March 2021 .....

prof. dr hab. Wojciech Niemiro

*Auxiliary Supervisor's declaration*
This dissertation is ready to be reviewed.

2 March 2021 .....

dr John Noble

# Acknowledgements

When I read a math book, I always like to start with the *acknowledgements* section. It is not only because this section is usually placed at the very beginning. I think that it tells a lot about its author's writing style and it reminds me that this author is (or was) a real person that put his time and heart into writing what I am about to read. Besides, acknowledgements can be pretty witty and humorous – a trait which does not always accompany you on the pages that follow. I simply could not miss an opportunity to write the *acknowledgements* section myself, and here it is.

Family comes first. I can always count on their support and this is wonderful. I thank my father for expressing the right amount of impatience in the creation of this dissertation and my mother on expressing an equal amount of understanding. Also I would like to thank my brother for several useful tricks that enhance concentration – I estimate that his help sped up the creation of this work by 5.19 weeks and you can take my word for it, I am a statistician after all.

I am grateful to my much more experienced colleagues from the statistical department (and its open neighbourhood). John Noble for being the most patient and understanding guide in this journey, long hours of discussions and going out for a pizza together from time to time. Wojciech Niemiro, for introducing me to the beautiful world of Bayesian Nonparametrics and, when the time was right, for a couple of substantial, wise words. Błażej Miasojedow for always being keen on answering my mathematical questions, even though sometimes his answers generated even more questions on my part. Przemek Biecek for his meaningful support and sharing the belief that explaining is important. Tomasz Rychlik for inviting me to give a presentation at the Institute of Mathematics (PAN) and for making some extremely inspiring comment afterwards.

Many thanks to Julyan Arbel for inviting me for a very pleasant stay in charming Grenoble. His interest in my research was a very important (and very needed at the time) incentive for future work. This visit, though short, left me with lots of good memories, for which I am grateful to the whole INRIA Mistis Team.

Friends from *Delta* monthly editorial team, for countless hours of inspiring discussions (not counting twice as much time of idle talk, which was also fun) and 'pulling this cart' together. It is certainly worth it, especially in your company. I also thank the creator of *Delta*, Marek Kordos, and his wife Krystyna for making the room 4020 (*Delta* headquarters) a second home to me.

To those of my colleagues from the faculty that shared with me the ups and downs of being a PhD student. Miś, Zosia, Grzesiek, Rafał, Mateusz, Kamila – thank you for all the laugh we shared and the feeling that we can always count on each other. Besides, I would like to thank the community of young (in hearts, regardless of age) statisticians, with particular attention paid to Agnieszka, Gosia, Masza, Krzysiek and Mariusz, for those great long evenings in Piekiełko and its surroundings during annual Będlewo conferences.

It is hard to overestimate the teacher's role in the development of scientific interest in a young man. I was lucky enough to have the best. Long time ago Krzysztof Zieleniewski found a mathematical spark in me and skillfully transformed it into a passion. Afterwards Wiktor Bartol and Jerzy Bednarczuk made sure that this passion is properly directed. I benefited a lot from their fantastic (and surely hard) didactic work.

Finally, there is Daria, my beloved wife. I guess she is the only person in the world (apart from myself) that is as glad that this dissertation has already been finished as I am. Throughout my PhD studies we went the road from being a couple, then betrothed, a married couple and parents. Yes, those were long studies, but it is not the point of this paragraph. The thing is that at each step she was infinitely supporting and patient, and for that – and just being herself – she has my twice-as-infinite gratitude and love.

# Contents

# Chapter 1

# Introduction

*The book* [of Nature] *is written in the mathematical language,* (...)

Galileo, 1610

This famous statement, quoted after Kline (1990), is often rephrased as *Mathematics is the language of Science.* I like to think about Mathematical Statistics in a similar vein – as the *language of Experimental Science.* Whether it is sociological research, drug development or search for elemental particles, results are established by analysing data based on principals of Mathematical Statistics. This language, or at least its basics, are understood by the whole scientific community and its development, together with a precise formulation of its limitations, is one of many ways in which Mathematics can contribute to the better understanding of the world around us.

As many languages do, Mathematical Statistics has its *dialects*, by which I mean different approaches to performing statistical inference. Among the most important there are two – *classical* and *Bayesian.* They differ in the treatment of the *parameter*, which guides the probabilistic mechanism in which data is assumed to be generated. In classical statistics this parameter is consider to be unknown, but fixed by Nature, whereas Bayesian paradigm assigns to it a probability distribution called the *prior distribution*, which is then updated to the *posterior distribution.* It is a basis for further inference, like the predictive distribution.

The posterior distribution is obtained by the usage of the Bayes Theorem. Even since high school the students should be familiar with its discrete form:

$$\mathbb{P}(B_i \,|\, A) = \frac{\mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}. \tag{1.1}$$

Thomas Bayes (1702–1761) established a related formula in a special case. He stated and solved the following problem. '*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named'. This comes from his posthumously published paper *An essay towards solving a problem in the*

*Doctrine of Chances* (Bayes, 1763). In a modern, *Bayesian* language we could translate it in a following way: compute the posterior probability of the parameter of the Bernoulli distribution given an independent sample from that distribution. The prior distribution is implicitly assumed to be uniform on $[0, 1]$. Hence he solved the continuous form of (1.1) in this case, as Fienberg et al. (2006) points out. The Bayesian approach is therefore appropriate nomenclature, since Bayes was the first to apply this form of reasoning to this class of problems. It was put on a solid mathematical footing by Pierre Simon Laplace (1749–1827).

The classical theory of statistics started with the works of Ronald Fisher (1890–1962). By introducing the notion of a parameter, Fisher revolutionised the way the scientific world thought about inferring from the data. At the time, Bayesian inference was not an appropriate tool as it was applicable to a very restricted modelling situations in which the computation of the posterior distribution is doable. This changed with the advent of high-performance computer hardware and sophisticated Markov Chain Monte Carlo methods, that allow to approximately sample from the posterior distribution. The latter were theoretically known to the statisticians since the 1970's, but they started to use them when it could be carried out on personal computers (Hjort et al., 2010).

This computational breakthrough also gave the incentive for the development of *nonparametric Bayes*. The term *nonparametric* is a bit misleading, as the parameter space is still present. This time though it is infinite-dimensional, contrary to the traditional, Bayesian setting. Examples of such parameter spaces include function spaces or spaces of spaces of probability distributions. The beginning of this non-parametric chapter in the Bayesian history is often identified with two seminal papers: Ferguson (1973) and Doksum (1974). The latter introduces neutral-to-the-right processes and the former describes a true celebrity in the non-parametric Bayesian world, namely the *Dirichlet process*. It is a random measure such that its finite dimensional distributions follow the Dirichlet distribution. Soon after this, Antoniak (1974) suggested how to use this construction to create a model where data does not necessarily come from a discrete distribution.

As easily anticipated, with great parameter space come great challenges and difficulties. Rather quickly after admitting infinitely dimensional spaces into the Bayesian world come examples in which the Bayesian inference does not behave as desired. Standard Bayesian procedures can be proved to be consistent, i.e. as the amount of information produced by a specific parameter value increases, the posterior distribution gets concentrated on this parameter value. One of the fundamental results in this regard for finite-dimensional parameter space is Doob's theorem, where the consistency is established using the theory of martingales. It is not always the case with the non-parametric methods. First counterexamples in infinitely dimensional spaces were given in Diaconis and Freedman (1986a) and Diaconis and Freedman (1986b).

These examples of inconsistency were rather theoretically involved and did not relate to any situation that could be encountered by a practitioner. However, since that time mathematicians constructed examples of non-parametric Bayesian inconsistency in situations

much closer to the hearts of a practice-oriented statisticians. One of them is presented in Miller and Harrison (2014) and concerns the inconsistency of the posterior for the number of clusters in non-parameteric Bayesian Mixture Models. These are introduced in more detail in Section 1.2, for now let us just mention that those models can be used for the inference about the clustering structure and in the aforementioned article the authors show a large class of situations in which the posterior probability of the number of clusters does not concentrate on the number of clusters in the mixture, from which the data was sampled from (Miller and Harrison, 2014, Corollary 1).

The practitioners often need some of summaries of the posterior distribution. In the need of a specific estimate of a parameter, the Bayesians turn to the decision theory, according to which they should choose the parameter value that minimises the posterior expected value of the risk function; the latter is chosen by the investigator. If the 0-1 loss function is chosen and the parameter space is discrete, the Bayes estimator is simply the mode of the posterior distribution, also called the maximum a posteriori probability estimator (MAP). This approach can be fruitfully applied even when the posterior probability is not continuous (in which case in general the MAP estimator is not a Bayesian estimator). One of the most prominent examples is the LASSO method for choosing the right coefficients in linear regression (Tibshirani, 1996); this estimator can be treated as the MAP estimator when the prior on the linear coefficients is the Laplace distribution.

In the discussion to their article, Miller and Harrison point out that the behaviour of the MAP (or other Bayesian) estimator is outside the scope of their analysis. This was the starting point of my research, aimed at investigation of the properties of this MAP clustering. In the course of my research I slightly altered the question – instead of the consistency analysis of the MAP estimator for the number of clusters I begun to investigate the number of clusters of the MAP estimator for the whole clustering structure, i.e. partition of data insto clusters. There were two reasons for that: firstly it seemed more computationally tractable and secondly it felt at least as important from the practitioner's point of view, since in the end of most of the cluster analysis examples we are interested in the clustering itself, not only the true number of groups. I believe that the results obtained in that research contribute, even if in a very restricted setting, to the better understanding of the properties of Bayesian Nonparametric methods for cluster analysis and hopefully they constitute an $\varepsilon$-brick to the pile of mathematical statistics.

## 1.1 Organisation of the Dissertation

In the remaining part of Chapter 1 we introduce the main mathematical notions, necessary for the rest of the dissertation, such as Bayesian Mixture Models (BMMs), the conjugate exponential model within clusters, the Maximum A Posteriori clustering. Chapter 2 consists of two important sections. In Section 2.1 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*). In Section 2.2 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 $\Delta_P$ function (depending also on the specification of the exponential conjugate model), since it is a difference of two functions that increase their values whenever two chunks of the partitions are merged. 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 the MAP clusterings of the independent sample from $P$ are related to the partitions of the observation space that maximise the $\Delta_P$ 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 fixed covariance Gaussian BMM (we later call a Normal-Normal BMM). These findings are presented in details in Chapter 3. 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, where we put a prior probability (Inverse-Wishart) on the within cluster covariance structure as well. At the same time, we observed some undesired behaviour of the $\Delta_P$ 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 $\Delta_P$ score. This is why in Chapter 4 an adjusted Normal-Inverse-Wishart model is considered, 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 Chapter 3. Finally, in this case as a limit we obtain a family of $\Delta_P$ functions that depend on the linear coefficient in the concentration parameter. We can translate this $\Delta_P$ functions to their empirical counterparts and hence obtain a convenient family of score function the measure the performance of data clustering. This 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 Chapter 4.

## 1.2 General Framework for BMMs

In this section we introduce the main object of our analysis, which is the Bayesian Mixture Model. Before stating the precise definition, we give an informal introduction. A *mixture* of probability distributions is their convex combination, e.g. $\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}(-1,1)+\frac{1}{6}\mathcal{N}(1,1)$ is a mixture of three normal distributions. In *mixture models* we assume that the data comes from an unknown mixture of distributions. The uncertainty is twofold: we do not know neither the mixture frequencies nor the component distributions. However, usually we assume that the latter come from some parametrised family of distributions, but we do not know the underlying parameters. The approach of this dissertation is Bayesian. We put a prior distribution both on the mixture frequencies (which is a probability distribution on the multi-dimensional simplex) and independent and identical priors on the parameters of the component distribution. Below we show a simple example of a Bayesian Mixture Model, which is a modification of the mixture of three normal distributions stated earlier that accounts for the uncertainty. Here $\mathrm{Dir}(\cdot)$ means the Dirichlet distribution.

$$\begin{aligned} \boldsymbol{p} = (p_1, p_2, p_3) &\sim \mathrm{Dir}(1,1,1) \\ \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) &\overset{\text{iid}}{\sim} \mathcal{N}(0,10) \\ \boldsymbol{x} = (x_1, \ldots, x_n) \,|\, \boldsymbol{p}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} p_1\mathcal{N}(\theta_1,1) + p_2\mathcal{N}(\theta_2,1) + p_3\mathcal{N}(\theta_3,1). \end{aligned} \tag{1.2}$$

It is probably a good moment to point out some notational convention. A careful (and, perhaps, a little pedantic) reader would note that the first two lines of (1.2) do not specify the joint distribution of $\boldsymbol{p}, \boldsymbol{\theta}$, only their marginal distributions. However it is popular in Bayesian literature to implicitly assume that the distribution of every new variable is specified conditionally on the variables previously introduced and if some of those are omitted in the description, this simply implies a relevant conditional independence statement. In case of (1.2) we implicitly assume that $\boldsymbol{p}$ and $\boldsymbol{\theta}$ are independent.

We are now ready to state the formal definition of the Bayesian Mixture Model in full generality. 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 $G_\theta$ has a density $g_\theta$ with respect to the Lebesgue measure. Those are the *component* measures, responsible for randomness within clusters. Consider a prior distribution $\vartheta$ on $\Theta$ (we will call it the *base* measure, defining how the parameters of the components are spread). Let $\pi$ 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

$$\begin{aligned} \boldsymbol{p} = (p_i)_{i=1}^m &\sim \pi \\ \boldsymbol{\theta} = (\theta_i)_{i=1}^m &\overset{\text{iid}}{\sim} \vartheta \\ \boldsymbol{x} = (x_1, \ldots, x_n) \,|\, \boldsymbol{p}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \textstyle\sum_{i=1}^m p_i G_{\theta_i}. \end{aligned} \tag{1.3}$$

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 (1.3) can be used to model data clustering; clusters are defined by deciding which distribution $G_{\theta_i}$ generated a given data point. To avoid confusion in the cluster assignment, from now on we assume that the base measure is nonatomic. In order to formally define the clusters, we need to rewrite (1.3) as

$$\begin{aligned} \boldsymbol{p} = (p_i)_{i=1}^m &\sim \pi \\ \boldsymbol{\theta} = (\theta_i)_{i=1}^m &\overset{\text{iid}}{\sim} \vartheta \\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \,|\, \boldsymbol{p}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \textstyle\sum_{i=1}^m p_i \delta_{\theta_i} \\ x_i \,|\, \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\phi} &\sim G_{\phi_i} \quad \text{for all } i \leq n. \end{aligned} \tag{1.4}$$

Then two observations with indices $i, j \leq n$ are in the same cluster if and only if $\phi_i = \phi_j$. In this way the distribution $\pi$ on the $m$-dimensional simplex *generates* a probability distribution $\mathcal{P}_{\pi,n}$ on the partitions of the index set $\{1, \ldots, n\}$ into at most $m$ subsets. We will use the notation $[n] := \{1, \ldots, n\}$, which is a popular convention in the literature. Note that the distribution of $\boldsymbol{\phi}$ is invariant to permutation of coordinates which implies that for any partition $\mathcal{I}$ of $[n]$ the probability weight $\mathcal{P}_{\pi,n}(\mathcal{I})$ depends only on the block sizes of $\mathcal{I}$. Following Pitman (2002, Section 2.1) we call such distributions on the space of partitions of $[n]$ *exchangeable*.

Let $\mathcal{P}_{\pi,n}$ be the exchangeable probability distribution on the space of partitions of $[n]$, generated by $\pi$. We can formulate (1.3) as follows: firstly we generate the partition of observations into clusters, 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 $G_\theta$. 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

$$g_{\vartheta,k}(u_1, \ldots, u_k) := \int_\Theta \prod_{i=1}^k g_\theta(u_i) \mathrm{d}\vartheta(\theta). \tag{1.5}$$

Now, (1.3) is equivalent to

$$\begin{aligned} \mathcal{I} &\sim \mathcal{P}_{\pi,n} \\ \boldsymbol{x}_I := (x_i)_{i \in I} \,|\, \mathcal{I} &\sim G_{\vartheta,|I|} \quad \text{for all } I \in \mathcal{I} \end{aligned} \tag{1.6}$$

We stress the fact that the (implicitly) independent sampling on 'the lowest' level of (1.6) relates to the independence between clusters (conditioned on the random partition); within one cluster the observations are (marginally) dependent. Using the within cluster conditional independence, we can write the density of $\boldsymbol{x}$ conditionally on $\mathcal{I}$:

$$g_{\vartheta,n}(\boldsymbol{x} \,|\, \mathcal{I}) := \prod_{I \in \mathcal{I}} g_{\vartheta,|I|}(\boldsymbol{x}_I). \tag{1.7}$$

Finally, for further convenience, let

$$Q(\boldsymbol{x}, \mathcal{I}) = \mathcal{P}_{\pi,n}(\mathcal{I}) \cdot g_{\vartheta,n}(\boldsymbol{x} \,|\, \mathcal{I}) \tag{1.8}$$

be the joint density of the partition and the observation. By Bayes rule, the expression in (1.8) 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 (1.8) 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.

Such a simplification does lead to loss of information, so does any point summary of the posterior distribution.

**Definition 1.1.** 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 (1.6) is any partition $\hat{\mathcal{I}}$ of $[n]$ that maximises $Q(\boldsymbol{x}, \mathcal{I})$ given by (1.8). In other words, the set of the MAP partitions is given by $\operatorname{argmax}_{\mathcal{I}} Q(\boldsymbol{x}, \mathcal{I})$.

As far as the applications are concerned, it may seem more convenient to speak about the MAP partition of the data set $\{x_1, \ldots, x_n\}$, not the set of indices $[n]$. However, the transition from the latter to the former is straightforward, and with our formulation (1.6) of the Bayesian Mixture Model the definition above is more appropriate.

## 1.3 The stick-breaking construction

In this dissertation we are concerned about the situation in which the number of mixture components is not bounded a priori. This is implied by the assumption that the distribution $\pi$ of the components' probability weights has a full support on the infinitely dimensional simplex $\triangle^\infty$. One of the most popular constructions of such probability distributions is the *stick-breaking construction*. It requires a sequence of independent random variables $V_i \in (0,1)$, $i \in \mathbb{N}$ that serve as the stick-breaking proportions. For the first probability $p_1$ we break the segment $[0,1]$ in proportion $V_1 : (1 - V_1)$ and take the left part. For the second probability we take the remaining part and break it in proportion $V_2 : (1 - V_2)$ and take the left part, and so on. Formally this relationship is described by

$$\begin{cases} p_1 = V_1, \\ p_k = V_k \prod_{i=1}^{k-1} (1 - V_i) \quad \text{for } k > 1, \end{cases} \tag{1.9}$$

and we let $\pi$ be the distribution of $\boldsymbol{p} = (p_1, p_2, \ldots)$. Note that for this definition to be valid, we need to ensure that $\sum_{i=1}^{\infty} p_i \overset{\text{a.s.}}{=} 1$. Since

$$\sum_{i=1}^{k} p_i = 1 - \prod_{i=1}^{k} (1 - V_i) \tag{1.10}$$

the condition $\sum_{i=1}^{\infty} p_i \overset{\text{a.s.}}{=} 1$ is equivalent to

$$\sum_{i=1}^{\infty} \log(1 - V_i) \overset{\text{a.s.}}{=} -\infty. \tag{1.11}$$

A prominent example of random variables $V_1, V_2, \ldots$ that satisfy (1.11) is a family of independent random variables such that

$$V_i \sim \text{Beta}(1 - \beta, \alpha + i\beta) \text{ for some } \alpha > 0 \text{ and } 0 \leq \beta < 1. \tag{1.12}$$

The proof that (1.11) is satisfied in this case is straightforward, but rather technical and hence often omitted in the literature. For the completeness of this dissertation we include it in the Appendix, see Lemma A.1 for details.

With (1.12), the distribution of the random measure $\sum_{i=1}^{\infty} p_i \delta_{\theta_i}$, where $\theta_i \,|\, \boldsymbol{p} \overset{\text{iid}}{\sim} \vartheta$, is the celebrated *Pitman-Yor process* (Pitman and Yor, 1997), which is a generalisation of the even more celebrated Dirichlet Process (Ferguson, 1973) – the latter being obtained by taking $\beta = 0$ in (1.11). The idea of the Dirichlet Process and the proof of its existence is considered the beginning of the Bayesian Nonparametrics and hence it seems only appropriate to devote a whole subsection for explaining the fundamentals of these notions.

### 1.3.1 The Dirichlet and the Pitman-Yor processes

The definition of the Dirichlet Process given in Ferguson (1973) is rather abstract. The Dirichlet Process on $\mathcal{X}$ with parameters $\alpha > 0$ (real number) and $\nu$ (a probability distribution on $\mathcal{X}$) is defined as a probability measure on the space of all probability distributions on $\mathcal{X}$ such that if $H$ is a random probability on $\mathcal{X}$ that is distributed according to this measure then for any finite measurable partition $(A_1, \ldots, A_k)$ of $\mathcal{X}$, the random vector $\big(H(A_1), \ldots, H(A_k)\big)$ has the Dirichlet distribution with parameters $\alpha\nu(A_1), \ldots, \alpha\nu(A_k)$. The existence of such random measure can be established by an application of the Kolmogorov extension theorem, which is not completely straightforward and requires attention to some measure-theoretical details; see Hjort et al. (2010, Chapter 2.2). If $H$ is distributed according to the Dirichlet Process with parameters $\alpha$ and $\nu$, we write $H \sim \text{DP}(\alpha, \nu)$.

Much more constructive proof of existence was given by Blackwell et al. (1973), where the Dirichlet Process is obtained using the *Generalised Polya Urn Scheme*. The scheme works as follows: firstly we sample $X_1 \sim \nu$ and then we let $X_{n+1} \,|\, X_1, \ldots X_n \sim \frac{\nu_n}{\alpha+n}$, where $\nu_n = \alpha\nu + \sum_{i=1}^{n} \delta_{X_i}$ ($\delta_x$ being the Dirac measure concentrated at $x$; note that $\alpha + n$ is the normalising constant of $\nu_n$). In Blackwell et al. (1973), the Dirichlet Process is proved to be the distribution of the almost sure limit in distribution of $\frac{\nu_n}{\alpha+n}$. Also a reverse statement is shown: if $H \sim \text{DP}(\alpha, \nu)$ and $X_1, \ldots, X_n \,|\, H \overset{\text{iid}}{\sim} H$, then $X_1 \sim \nu$ and $X_{n+1} \,|\, X_1, \ldots X_n \sim \frac{\nu_n}{\alpha+n}$. This nice conditional structure of independent samples from a realisation of the Dirichlet Process makes various Markov Chain Monte Carlo algorithms efficient, which enables an inference from this conceptually complicated model – this is one of the reasons of its popularity.

Sethuraman (1994) formally proved that the distribution of a random measure $\sum_{i=1}^{\infty} p_i \delta_{\theta_i}$, where $p_i$ are defined by (1.9) and (1.12) with $\beta = 0$ and $\theta_i \overset{\text{iid}}{\sim} \vartheta$, is the Dirichlet Process with parameters $\alpha$ and $\vartheta$. Hence (1.4) can be rewritten as

$$\begin{aligned} H &\sim \mathrm{DP}(\alpha, \vartheta) \\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \,|\, H &\overset{\text{iid}}{\sim} H \\ x_i \,|\, \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\phi} &\sim G_{\phi_i} \quad \text{independently for all } i \leq n. \end{aligned} \tag{1.13}$$

This is the *Dirichlet Process mixture model*, introduced by Antoniak (1974). The result of Blackwell et al. (1973) allows us to specify the induced probability $\mathcal{P}_{\pi,n}$ on the space of partitions of $[n]$. This is so called *Chinese Restaurant Process*, the name coined in Aldous (1985). 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 $\alpha$ 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 $\alpha$ 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 $[n]$:

$$\mathcal{P}_{\pi,n}(\mathcal{I}) = \frac{\alpha^{|\mathcal{I}|}}{\alpha^{(n)}} \prod_{I \in \mathcal{I}} (|I| - 1)!, \tag{1.14}$$

where $\alpha^{(n)} = \alpha(\alpha+1)\ldots(\alpha+n-1)$. The Chinese Restaurant Process can be generalised in the following way: the $n$-th customer chooses a table used by $n_i > 0$ customers with probability proportional to $n_i - \beta$ and she chooses the new table with probability proportional to $\alpha + K\beta$, where $K$ is the number of already occupied tables. This translates to the following probability weights of partitions of $[n]$:

$$\mathcal{P}_{\pi,n}(\mathcal{I}) = \frac{\alpha^{(|\mathcal{I}| \nearrow \beta)}}{\alpha^{(n)}} \prod_{I \in \mathcal{I}} (1 - \beta)^{(|I|-1)}, \tag{1.15}$$

where $\alpha^{(n \nearrow d)} = \alpha(\alpha + d)\ldots(\alpha + (n-1)d)$. In the seminal paper by Pitman and Yor (1997) it was proved that (1.15) corresponds to the stick-breaking construction with 'stick-breaking proportions' given by (1.12). A nice and elementary proof of this relationship can be also found in Lawless and Arbel (2019).

## 1.4 Conjugate exponential families

In Section 1.3 we dealt with the possible prior distribution on the mixture weights in Bayesian Mixture Models. In this section we briefly present a natural and computationally convenient candidates for distributions $\nu$ and $G_\theta$ (the base and the component measures), namely conjugate exponential families.

We start with a short reminder of the *exponential families*. This fundamental class of statistical models was developed independently by Darmois (1935), Koopman (1936) and

Pitman (1936) (a genealogical note: the last one being the father of Jim Pitman, responsible, together with Mihael Perman and Marc Yor, for the Pitman–Yor process, described in Section 1.3.1). 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 $\theta$ the probability $G_\theta$ has the following density with respect to the Lebesgue measure:

$$g_\theta(x) = h(x) \cdot \exp\left\{T(x)^\top \eta(\theta) - \mathsf{B}(\theta)\right\}, \tag{1.16}$$

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. Here we treat elements of $\mathbb{R}^k$, $k = 1, 2, \ldots$ as column vectors and for $v \in \mathbb{R}^k$ by $v^\top$ we denote its transpose, so that $T(x)^\top \eta(\theta)$ in (1.16) is simply the standard scalar product of $T(x)$ and $\eta(\theta)$ in $\mathbb{R}^p$. Clearly $\mathsf{B}$ in (1.16) is implicitly defined as

$$\mathsf{B}(\theta) = \log \int_{\mathcal{X}} h(x) \cdot \exp\left\{T(x)^\top \eta(\theta)\right\} \mathrm{d}x \tag{1.17}$$

so that (1.16) is indeed a density function and integrates to 1.

If we let the model be indexed by $\eta = \eta(\theta)$ rather than $\theta$ 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

$$g'_\eta(x) = h(x) \cdot \exp\left\{T(x)^\top \eta - \mathsf{A}(\eta)\right\}, \tag{1.18}$$

where

$$\mathsf{A}(\eta) = \log \int_{\mathcal{X}} h(x) \cdot \exp\left\{T(x)^\top \eta\right\} \mathrm{d}x \tag{1.19}$$

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

$$\mathcal{E} = \{\eta \in \mathbb{R}^p \colon \mathsf{A}(\eta) < \infty\} \tag{1.20}$$

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 $g_\theta$ is given by (1.16)) when the corresponding canonical form is regular and $\theta \colon \Theta \to \mathcal{E}$ is a bijection. We point out a minor abuse of notation, as $\theta$ and $\eta$ are treated both as parameters and transformations between parameter spaces. This convention is standard in the literature and rarely leads to misunderstandings.

The following result is a standard property of exponential families.

**Theorem 1.2** (parts (a) and (b) of Theorem 1.6.3 in Bickel and Doksum (2015))**.** *Consider a canonical exponential family, where the densities are given by (1.18), with the corresponding log-partition function $\mathsf{A}(\eta)$ and the natural parameter space $\mathcal{E}$. Then*

*(a) $\mathcal{E}$ is a convex set,*

*(b) $\mathsf{A} \colon \mathcal{E} \to \mathbb{R}$ is a convex function.*

Note that we can treat (1.19) as a definition of a function on $\mathbb{R}^p$ with values in $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. In this way $\mathsf{A}$ remains a convex function in the sense presented in Rockafellar (1970) (Section 4). Assuming the terminology from Rockafellar (1970), the natural parameter space (defined by (1.20)) becomes an *essential domain* of the log-partition function $\mathsf{A}$.

In this dissertation we will usually rely on the *strict* convexity of the log-partition function $\mathsf{A}$. Another standard result in the theory of exponential families is that the strict convexity is implied by the *identifiability* of the canonical parameter, i.e. the fact that different canonical parameters define different distributions (Bickel and Doksum, 2015, Theorem 1.6.4).

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 $\theta$ comes from this family and the likelihood is given by (1.16), then the posterior distribution $\theta \,|\, \boldsymbol{x}$ also belongs to this family.

Suppose that in (1.16) we can write $\mathsf{B}(\theta)$ as $\mathsf{B}(\theta) = \boldsymbol{a}^\top \mathbf{B}(\theta)$ where $\boldsymbol{a} \in \mathbb{R}^q$ and $\mathbf{B}(\theta) = [\mathsf{B}_1(\theta), \ldots, \mathsf{B}_q(\theta)]^\top$. Consider a canonical exponential family on $\Theta$, where the densities are given by

$$\gamma_{\tau,\zeta}(\theta) := \psi(\theta) \cdot \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} \tau \\ \zeta \end{bmatrix} - \mathsf{C}(\tau, \zeta) \right\}, \tag{1.21}$$

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, given by

$$\mathsf{C}(\tau, \zeta) := \log\left( \int_\Theta \psi(\theta) \cdot \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} \tau \\ \zeta \end{bmatrix} \right\} \mathrm{d}\theta \right) \tag{1.22}$$

Let $\Omega$ be the natural (hyper)parameter space, i.e.

$$\Omega := \{(\tau, \zeta) \in \mathbb{R}^p \times \mathbb{R}^q \colon \mathsf{C}(\tau, \zeta) < \infty\} \tag{1.23}$$

(we assume that it is non-empty). It follows that if $\theta \sim \vartheta$, where $\vartheta$ has density $\gamma_{\tau_0,\zeta_0}$ for some $(\tau_0, \zeta_0) \in \Omega$ and $\boldsymbol{x} = (x_1, \ldots, x_k) \,|\, \theta \overset{\text{iid}}{\sim} g_\theta$ then the joint density of $(\theta, \boldsymbol{x})$ is

$$\gamma_{\tau_0,\zeta_0}(\theta) \prod_{i=1}^k g_\theta(x_i) = \psi(\theta) \prod_{i=1}^k h(x_i) \cdot \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} \tau_0 + \sum_{i=1}^k T(x_i) \\ \zeta_0 + k\boldsymbol{a} \end{bmatrix} - \mathsf{C}(\tau_0, \zeta_0) \right\}. \tag{1.24}$$

The conditional density of $\theta \,|\, \boldsymbol{x}$ is proportional to (1.24) as a function of $\theta$. Comparing this with (1.21) we see that $\theta \,|\, \boldsymbol{x} \sim \gamma_{\tau_{\boldsymbol{x}},\zeta_k}$, where $\tau_{\boldsymbol{x}} := \tau_0 + \sum_{i=1}^k T(x_i)$ and $\zeta_k := \zeta_0 + k\boldsymbol{a}$, i.e.

$$\theta \,|\, \boldsymbol{x} \sim \psi(\theta) \cdot \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} \tau_0 + \sum_{i=1}^k T(x_i) \\ \zeta_0 + k\boldsymbol{a} \end{bmatrix} - \mathsf{C}\Big(\tau_0 + \sum_{i=1}^k T(x_i), \zeta_0 + k\boldsymbol{a}\Big) \right\}. \tag{1.25}$$

As the marginal density of $\boldsymbol{x}$ is the quotient of the joint density of $(\theta, \boldsymbol{x})$ and the conditional density of $\theta \,|\, \boldsymbol{x}$, by dividing (1.24) by (1.25) we get that

$$\boldsymbol{x} \sim g_{\vartheta,k}(\boldsymbol{x}) = \prod_{i=1}^k h(x_i) \cdot \exp\{\mathsf{C}(\tau_{\boldsymbol{x}}, \zeta_k) - \mathsf{C}(\tau_0, \zeta_0)\} \tag{1.26}$$

**Note 1.3.** This definitions of conjugate exponential family is the multi-dimensional analogue of the definition usually given in standard texts, e.g. Bickel and Doksum, Section 1.6.5, which corresponds to the definition above with $q = 1$ and $\boldsymbol{a} = 1$. The multi-dimensional version is needed in Section 1.4.1 to deal with multivariate normal models.

**Convexity assumption.** From Theorem 1.2 it follows that the function $\boldsymbol{a}^\top \mathbf{B}\big(\theta(\eta)\big) = \mathsf{B}\big(\theta(\eta)\big) = \mathsf{A}(\eta)$ is a convex function on $\mathcal{E}$. In Section 2.2 we will assume that also

$$\text{for any } (\tau_0, \zeta_0) \in \Omega \text{ the function } {\zeta_0}^\top \mathbf{B}\big(\theta(\eta)\big) \text{ is a convex function on } \mathcal{E}. \tag{1.27}$$

We call this assumption a *convexity assumption*; it is satisfied by all multivariate conjugate Normal models presented in Section 1.4.1. Moreover, when $q = \boldsymbol{a} = 1$, the convexity assumption is equivalent to $\zeta_0 > 0$. When $\psi(\theta)$ is a constant, this inequality follows from the definition of $\Omega$. This implication is proved in Diaconis and Ylvisaker, 1979, Theorem 1.

**Definition 1.4.** *Canonical Exponential Family Bayesian Mixture Model* is a Bayesian Mixture Model in which the component density is given by (1.16) and the base density is (1.21) for some $(\tau_0, \zeta_0) \in \Omega$.

### 1.4.1 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 $G_\theta$ are multivariate Normal. This corresponds to the data being normally distributed within clusters, which is a rather standard assumption. The cluster location, i.e. the mean of respective normal distribution, is also assumed to be normally distributed. Here we consider three possibilities concerning the covariance structure within clusters: it can be treated as known, unknown or known up to a scaling factor. If it is known, it is treated as the hyperparameter of the model, and hence the parameter space $\Theta$ is just the space of possible component means, which is $\mathbb{R}^d$, and $\vartheta$ is some fixed multivariate Normal distribution. When the component covariance is unknown, the parameter space becomes the product of $\mathbb{R}^d$ (for mean locations) and the space of $d$-dimensional positive definite matrices $\mathcal{S}_+^d$ (for covariance structures). The marginal distribution of the covariance parameter is the Inverse-Wishart distribution. In the case when the components covariance structure is known up to the scale factor, the marginal distribution of this factor is simply the Inverse-Gamma distribution. Since some of the results of this dissertation relate specifically to the Normal mixture model, here we present in detail relevant formulas.

**Notation.** We use two standard notations to denote the determinant of a square matrix $\Lambda$: $\det \Lambda$ and $|\Lambda|$. The latter may seem ambiguous as we also use the symbol $|\cdot|$ to denote the cardinality of a set and absolute value of a real number. However, the meaning of this symbol is always clear from the context.

**Notation.** To keep the notation precise, we introduce the following convention: if $\Sigma$ is a symmetric $d \times d$ matrix, then $\mathrm{diag}(\Sigma)$ is the diagonal of $\Sigma$, treated as $d$-dimensional vector, and $\mathrm{low}(\Sigma)$ is the 'lower triangular' part of $\Sigma$, treated as a $\frac{d(d-1)}{2}$ dimensional vector, whose $\left(\frac{(i-1)(i-2)}{2} + j\right)$-th coordinate is equal to $(i,j)$-th coefficient of $\Sigma$, where $i > j$. Whenever this notation occurs it may seem artificial but without it it is difficult to represent relevant models in the form of exponential families, given by (1.16).

## 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 $x \,|\, \mu \sim \mathcal{N}(\mu, \Sigma_0)$, where $\Sigma_0$ is known. The base measure is

$$\mu \quad \sim \quad \mathcal{N}(\mu_0, \Psi_0). \tag{1.28}$$

The hyperparameters are $\mu_0 \in \mathbb{R}^d$ and $\Psi_0, \Sigma_0 \in \mathcal{S}_+^d$. This prior is listed in Gelman et al. (2013) and it is a rather standard example of a conjugate Bayesian family. Clearly

$$\mathbb{E}\left(\mathbf{V}(x \,|\, \mu)\right) = \Sigma_0, \quad \mathbf{V}(\mathbb{E}\left(x \,|\, \mu\right)) = \mathbf{V}(\mu) = \Psi_0. \tag{1.29}$$

where $\mathbf{V}(\cdot)$ is the (conditional) covariance matrix. The conditional densities are given by

$$\begin{aligned} x \,|\, \mu \quad &\sim \quad (2\pi)^{-d/2}|\Sigma_0|^{-1/2} \exp\left\{-\tfrac{1}{2}(x-\mu)^\top \Sigma_0^{-1}(x-\mu)\right\} \\ \mu \quad &\sim \quad (2\pi)^{-d/2}|\Psi_0|^{-1/2} \exp\left\{-\tfrac{1}{2}(\mu-\mu_0)^\top \Psi_0^{-1}(\mu-\mu_0)\right\} \end{aligned} \tag{1.30}$$

The density of $x \,|\, \theta$ can be expressed as (1.16) by placing

$$\begin{aligned} h(x) = (2\pi)^{-d/2}|\Sigma_0|^{-1/2} \exp\left\{-\frac{1}{2}x^\top \Sigma_0^{-1} x\right\}, \quad T(x) = \Sigma_0^{-1}x, \quad \eta(\theta) = \mu, \\ \mathsf{B}(\theta) = \frac{1}{2}\mu^\top \Sigma_0^{-1}\mu \end{aligned} \tag{1.31}$$

We get the density of $\mu$ in (1.30) expressed as (1.21) by placing

$$\boldsymbol{a} = \begin{bmatrix} \mathrm{diag}(\Sigma_0^{-1}) \\ \mathrm{low}(\Sigma_0^{-1}) \end{bmatrix}, \quad \mathbf{B}(\theta) = \begin{bmatrix} \frac{1}{2}\mathrm{diag}(\mu\mu^\top) \\ \mathrm{low}(\mu\mu^\top) \end{bmatrix}, \quad \zeta_0 = \begin{bmatrix} \mathrm{diag}(\Psi_0^{-1}) \\ \mathrm{low}(\Psi_0^{-1}) \end{bmatrix}, \quad \tau_0 = \Psi_0^{-1}\mu_0 \tag{1.32}$$

and

$$\psi(\theta) = (2\pi)^{-d/2}, \quad \mathsf{C}(\tau_0, \zeta_0) = \frac{1}{2}\log|\Psi_0| + \frac{1}{2}{\mu_0}^\top \Psi_0^{-1}\mu_0. \tag{1.33}$$

Plugging (1.33) into (1.26) we get that the marginal density of $\boldsymbol{x} = (x_1, \ldots, x_k)$ in the Normal-Normal model is given by

$$g_{NN,k}(\boldsymbol{x}) = \frac{|\Psi_k|^{1/2}}{(2\pi)^{dk/2}|\Psi_0|^{1/2}|\Sigma_0|^{dk/2}} \exp\left\{-\frac{1}{2}W(\boldsymbol{x})\right\}, \tag{1.34}$$

where

$$\Psi_k = (\Psi_0^{-1} + k\Sigma_0^{-1})^{-1} \quad \text{and} \tag{1.35}$$

$$W(\boldsymbol{x}) = \sum_{i=1}^{k} (x_i - \mu_0)^\top \Sigma_0^{-1} (x_i - \mu_0) - k^2 \big(\overline{\boldsymbol{x}} - \mu_0\big)^\top \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} \big(\overline{\boldsymbol{x}} - \mu_0\big), \tag{1.36}$$

in which $\overline{\boldsymbol{x}}$ is the standard notation for the mean vector $\frac{1}{k}\sum_{i=1}^{k} x_i$. The detailed derivation of (1.34) from (1.26) can be found in Appendix A.1.1.

**Regularity.** The regularity of the Normal-Normal model is straightforward. It is already in its canonical form, and the parameter space $\Theta = \mathbb{R}^d$ clearly satisfies the definition (1.20) of the natural parameter space.

**Convexity assumption.** The proof that the convexity assumption (1.27) is satisfied for the Normal-Normal model is simple. Let us analyse (1.32). The function $\boldsymbol{a}^\top \mathbf{B}\big(\theta(\eta)\big)$ is convex on $\mathcal{E}$ since it is the log partition function. This is true for any $\Sigma_0 \in \mathcal{S}_+^d$. It is now clear that the function ${\zeta_0}^\top \mathbf{B}\big(\theta(\eta)\big)$ is convex as well, which follows from taking $\Sigma_0 = \Psi_0$.

### 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. This can be naturally interpreted as an open subset of $\mathbb{R}^p$, where $p = \frac{d(d-1)}{2} + d$. For $\theta = (\mu, \Lambda) \in \Theta$ the component distribution is $x \,|\, \theta \sim \mathcal{N}(\mu, \Lambda)$ and the base measure $\vartheta$ on $(\mu, \Lambda)$ is defined by the following conditional structure

$$\begin{aligned} \Lambda &\sim \mathcal{W}^{-1}(\nu_0 + d + 1, \nu_0 \Sigma_0) \\ \mu \,|\, \Lambda &\sim \mathcal{N}(\mu_0, \Lambda/\kappa_0), \end{aligned} \tag{1.37}$$

where $\mathcal{W}^{-1}$ is the Inverse-Wishart distribution. Here the hyperparameters are $\kappa_0, \nu_0 > 0$, $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0 \in \mathcal{S}_+^d$. This model is also listed in Gelman et al. (2013) with a slightly different parametrisation of the Inverse-Wishart distribution, but we made this modification to obtain

$$\begin{aligned} \mathbb{E}\left(\mathbf{V}(x \,|\, \mu, \Lambda)\right) &= \mathbb{E}\,\Lambda = \Sigma_0, \\ \mathbf{V}(\mathbb{E}\,(x \,|\, \mu, \Lambda)) &= \mathbf{V}(\mu) = \mathbb{E}\,\mathbf{V}(\mu \,|\, \Lambda) + \mathbf{V}\mathbb{E}\,(\mu \,|\, \Lambda) = \mathbb{E}\,\Lambda/\kappa_0 + \mathbf{V}(\mu_0) = \Sigma_0/\kappa_0, \end{aligned} \tag{1.38}$$

which is consistent with the Normal-Normal model, described earlier. It is also worth pointing out that the variance of the $(i, j)$-th conditional covariance is

$$\mathrm{Var}(\mathbf{V}(x \,|\, \mu, \Lambda)_{ij}) = \mathrm{Var}(\Lambda_{ij}) = \frac{(\nu_0 + 2)\sigma_{ij}^2 + \nu_0 \sigma_{ii}\sigma_{ij}}{(\nu_0 + 1)(\nu_0 - 2)}, \tag{1.39}$$

where $\sigma_{ij}$ are the coefficients of $\Sigma_0$ (for detailed derivation see e.g. Press (2005)). Hence it is clear that $\nu_0$ can be treated as 'concentration parameter' and the larger it is, the more the random covariance structure of the clusters is concentrated around $\Sigma_0$.

The conditional densities are given by

$$\begin{aligned} x \,|\, \mu, \Lambda &\sim (2\pi)^{-d/2}|\Lambda|^{-1/2} \exp\left\{-\tfrac{1}{2}(x-\mu)^\top \Lambda^{-1}(x-\mu)\right\} \\ \mu \,|\, \Lambda &\sim (2\pi)^{-d/2}\kappa_0^{d/2}|\Lambda|^{-1/2} \exp\left\{-\tfrac{1}{2}\kappa_0(\mu-\mu_0)^\top \Lambda^{-1}(\mu-\mu_0)\right\} \\ \Lambda &\sim \left(\tfrac{|\nu_0\Sigma_0|}{2^d}\right)^{\frac{\nu_0+d+1}{2}} \Gamma_d\left(\tfrac{\nu_0+d+1}{2}\right)^{-1} |\Lambda|^{-\frac{\nu_0+2d+2}{2}} \exp\left\{-\tfrac{1}{2}\nu_0 \mathrm{tr}(\Sigma_0\Lambda^{-1})\right\} \end{aligned} \tag{1.40}$$

The density of $x \,|\, \theta$ can be expressed as (1.16) by placing

$$\begin{gathered} h(x) = (2\pi)^{-d/2}, \quad T(x) = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(xx^\top) \\ -\mathrm{low}(xx^\top) \\ x \end{bmatrix}, \quad \eta(\theta) = \begin{bmatrix} \mathrm{diag}(\Lambda^{-1}) \\ \mathrm{low}(\Lambda^{-1}) \\ \Lambda^{-1}\mu \end{bmatrix}, \\ \mathsf{B}(\theta) = \frac{1}{2}\log|\Lambda| + \frac{1}{2}\mu^\top\Lambda^{-1}\mu \end{gathered} \tag{1.41}$$

where $\Gamma_d$ is the multivariate Gamma function given by

$$\Gamma_d(x) = \pi^{\frac{d(d-1)}{4}} \prod_{i=0}^{d-1} \Gamma\left(x - \frac{i}{2}\right). \tag{1.42}$$

We get (1.21) by placing

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{B}(\theta) = \begin{bmatrix} \frac{1}{2}\log|\Lambda| \\ \frac{1}{2}\mu^\top\Lambda^{-1}\mu \end{bmatrix}, \quad \zeta_0 = \begin{bmatrix} \nu_0 + 2d + 3 \\ \kappa_0 \end{bmatrix}, \quad \tau_0 = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\nu_0\Sigma_0 + \kappa_0\mu_0\mu_0{}^\top) \\ -\mathrm{low}(\nu_0\Sigma_0 + \kappa_0\mu_0\mu_0{}^\top) \\ \kappa_0\mu_0 \end{bmatrix} \tag{1.43}$$

and

$$\psi(\theta) = (2\pi)^{-d/2}, \quad \mathsf{C}(\tau_0, \zeta_0) = -\frac{d}{2}\log\kappa_0 - \frac{\nu_0+d+1}{2}\log\frac{|\nu_0\Sigma_0|}{2^d} + \log\Gamma_d\left(\frac{\nu_0+d+1}{2}\right). \tag{1.44}$$

By an application of (1.26) we get that in this case the marginal density of $\boldsymbol{x} = (x_1, \ldots, x_k)$ is given by

$$g_{NIW,k}(\boldsymbol{x}) = \frac{|\nu_0\Sigma_0|^{(\nu_0+d+1)/2}\kappa_0^{d/2}\Gamma_d\left(\frac{\nu_k+d+1}{2}\right)}{\pi^{dk/2}{\kappa_k}^{d/2}\Gamma_d\left(\frac{\nu_0+d+1}{2}\right)} \cdot \det\left(\Sigma(\boldsymbol{x})\right)^{-(\nu_k+d+1)/2}, \tag{1.45}$$

where $\Gamma_d$ is the multivariate Gamma function and

$$\nu_k = \nu_0 + k, \;\; \kappa_k = \kappa_0 + k \quad \text{and} \tag{1.46}$$

$$\Sigma(\boldsymbol{x}) = \nu_0\Sigma_0 + \sum_{i=1}^{k}(x_i - \overline{\boldsymbol{x}})(x_i - \overline{\boldsymbol{x}})^\top + \frac{\kappa_0 k}{\kappa_k}(\overline{\boldsymbol{x}} - \mu_0)(\overline{\boldsymbol{x}} - \mu_0)^\top. \tag{1.47}$$

This formula can be quickly deduced from Murphy (2007, eq. 266). We also give a detailed derivation of (1.45) from (1.26) in Appendix A.1.2.

**Regularity.** Clearly, $\eta\colon \Theta \to \mathbb{R}^{d(d+3)/2}$ is an injective function. We now show that $\eta(\Theta) = \mathcal{E}$. Suppose that $\eta_0 \in \mathcal{E}$, i.e.

$$\int_{\mathcal{X}} \exp\left\{-\frac{1}{2}(x-\mu_0)^\top \Lambda_0^{-1}(x-\mu_0)\right\} < \infty, \tag{1.48}$$

where the $d$-dimensional vector $\mu_0$ and $d\times d$ symmetric matrix $\Lambda_0$ are defined by the equality $\eta_0 = [\mathrm{diag}(\Lambda_0^{-1})^\top, \mathrm{low}(\Lambda_0^{-1})^\top, \Lambda_0^{-1}\mu_0{}^\top]^\top$. It is enough to show that $\Lambda_0$ (equivalently, $\Lambda_0^{-1}$) is positive definite. Suppose the contrary, i.e. that there exist $v \in \mathbb{R}^d$ such that $v^\top \Lambda_0^{-1} v \le 0$. Let $D = v^\perp \cap B(\mathbf{0}^d, 1)$ and let $T = \{\alpha v + \Lambda_0 w\colon \alpha \in \mathbb{R}, w \in D\}$ be an infinite cylinder. Then, for any $u \in T$, $u = \alpha v + \Lambda_0 w$:

$$u^\top \Lambda_0^{-1} u = (\alpha v + \Lambda_0 w)^\top \Lambda_0^{-1}(\alpha v + \Lambda_0 w) = \alpha^2 v^\top \Lambda_0^{-1} v + 2\alpha v^\top w + w^\top \Lambda_0 w \le \|\Lambda_0\|, \tag{1.49}$$

since $v^\top \Lambda_0^{-1} v \le 0$, $v^\top w = 0$ and $w^\top \Lambda_0 w \le \|w\| \cdot \|\Lambda_0 w\| \le \|\Lambda_0\|$. It follows that for the infinite cylinder $\mu_0 + T$, the value of the function under the integral in (1.48) is bounded from below by $\exp\{-\frac{1}{2}\|\Lambda_0\|\}$. This clearly contradicts (1.48) and the regularity follows.

**Convexity assumption.** By Theorem 1.2 the function $\mathsf{B}\big(\theta(\eta)\big) = \boldsymbol{a}^\top \mathbf{B}\big(\theta(\eta)\big)$ is a convex function of $\eta$. This does not imply (1.27) in a straightforward way. However, fix any $t > 0$ and suppose that the component distribution is $x \,|\, \theta \sim \mathcal{N}(\mu, t\Lambda)$. Then in (1.41) the formula for $\mathsf{B}(\theta)$ becomes $\mathsf{B}'(\theta) = \frac{d\log t}{2} + \frac{1}{2}\log|\Lambda| + \frac{1}{2t}\mu^\top \Lambda^{-1}\mu$ and we can leave the formula for $\eta(\theta)$ untouched by applying a relevant modification of $T(x)$, i.e. $T'(x) = T(x)/t$. Again by Theorem 1.2, the function $\mathsf{B}'\big(\theta(\eta)\big)$ is convex, and hence the function $\eta \mapsto \frac{1}{2}\log|\Lambda(\eta)| + \frac{1}{2t}\mu(\eta)^\top \Lambda(\eta)^{-1}\mu(\eta)$ is convex (since $\frac{t\log d}{2}$ is a constant). This implies that $[1, 1/t]\mathbf{B}\big(\theta(\eta)\big)$ is a convex function. As the choice of $t > 0$ was arbitrary, we get (1.27).

## Normal-Inverse-Gamma (NIG)

Note that although the Normal-Inverse-Wishart prior gives more flexibility in terms of the component covariances, it imposes some modelling restriction, namely the component covariance matrix $\Lambda$ and the covariance matrix of the component mean $\Lambda/\kappa_0$ are proportional random matrices. This is the reason for which here we also consider the Normal-Inverse-Gamma prior. We were not able to find any reference to it in the literature. It is not listed in Gelman et al. (2013) and in Murphy (2007, Chapter 6) only its 1-dimensional version is considered (which can be also treated as a one-dimensional version of the Normal-Inverse-Wishart model). It only allows a 1-parameter variation of the covariance function, but no restrictions are imposed on the within-group means, unlike the Normal-Inverse-Wishart prior.

In Normal-Inverse-Gamma model we assume that the base covariance matrix and the component covariance matrix are known up to some scaling factor $\lambda \sim \mathcal{G}^{-1}(\beta_0 + 1, \beta_0\gamma_0)$. Hence the parameter is $\theta = (\mu, \lambda)$, the parameter space is $\Theta = \mathbb{R}^d \times \mathbb{R}_+$, the component

distributions are $x \,|\, \mu, \lambda \sim \mathcal{N}(\mu, \lambda\Sigma_0)$ and the base measure on $(\mu, \lambda)$ is defined by the following conditional structure:

$$\begin{aligned} \lambda &\sim \mathcal{G}^{-1}(\beta_0 + 1, \beta_0\gamma_0) \\ \mu \,|\, \lambda &\sim \mathcal{N}(\mu_0, \lambda\Psi_0) \end{aligned} \tag{1.50}$$

Here the hyperparameters are $\beta_0, \gamma_0 > 0$, $\mu_0 \in \mathbb{R}^d$ and $\Psi_0, \Sigma_0 \in \mathcal{S}_+^d$. With this prior

$$\begin{aligned} \mathbb{E}\left(\mathbf{V}(x \,|\, \mu, \lambda)\right) &= \mathbb{E}\,\lambda\Sigma_0 = \gamma_0\Sigma_0, \\ \mathbf{V}(\mathbb{E}\left(x \,|\, \mu, \lambda\right)) &= \mathbf{V}(\mu) = \mathbb{E}\,\mathbf{V}(\mu \,|\, \lambda) + \mathbf{V}\mathbb{E}\left(\mu \,|\, \lambda\right) = \mathbb{E}\,\lambda\Psi_0 + \mathbf{V}(\mu_0) = \gamma_0\Psi_0. \end{aligned} \tag{1.51}$$

The conditional densities are given by

$$\begin{aligned} x \,|\, \mu, \lambda &\sim (2\pi)^{-d/2}|\Sigma_0|^{-1/2}\lambda^{-d/2}\exp\left\{-\tfrac{1}{2\lambda}(x-\mu)^\top\Sigma_0^{-1}(x-\mu)\right\} \\ \mu \,|\, \lambda &\sim (2\pi)^{-d/2}|\Psi_0|^{-1/2}\lambda^{-d/2}\exp\left\{-\tfrac{1}{2\lambda}(\mu-\mu_0)^\top\Psi_0^{-1}(\mu-\mu_0)\right\} \\ \lambda &\sim (\beta_0\gamma_0)^{\beta_0+1}\Gamma(\beta_0+1)^{-1}\lambda^{-(\beta_0+2)}\exp\left\{-\beta_0\gamma_0/\lambda\right\} \end{aligned} \tag{1.52}$$

The density of $x \,|\, \theta$ can be expressed as (1.16) by placing

$$\begin{gathered} h(x) \equiv (2\pi)^{-d/2}|\Sigma_0|^{-1/2}, \quad T(x) = \begin{bmatrix} -\frac{1}{2}x^\top\Sigma_0^{-1}x \\ \Sigma_0^{-1}x \end{bmatrix}, \quad \eta(\theta) = \begin{bmatrix} 1/\lambda \\ \mu/\lambda \end{bmatrix}, \\ \mathsf{B}(\theta) = \frac{d}{2}\log\lambda + \frac{1}{2}\mu^\top\Sigma_0^{-1}\mu/\lambda \end{gathered} \tag{1.53}$$

We get (1.21) by placing

$$\boldsymbol{a} = \begin{bmatrix} d/2 \\ \mathrm{diag}(\Sigma_0^{-1}) \\ \mathrm{low}(\Sigma_0^{-1}) \end{bmatrix}, \quad \mathbf{B}(\theta) = \begin{bmatrix} \log\lambda \\ \frac{1}{2}\mathrm{diag}(\mu\mu^\top)/\lambda \\ \mathrm{low}(\mu\mu^\top)/\lambda \end{bmatrix}, \quad \zeta_0 = \begin{bmatrix} \beta_0 + 2 \\ \mathrm{diag}(\Psi_0^{-1}) \\ \mathrm{low}(\Psi_0^{-1}) \end{bmatrix}, \quad \tau_0 = \begin{bmatrix} -\beta_0\gamma_0 - \frac{1}{2}\mu_0^\top\Psi_0^{-1}\mu_0 \\ \Psi_0^{-1}\mu_0 \end{bmatrix} \tag{1.54}$$

and

$$\psi(\theta) = (2\pi\lambda)^{-d/2}, \quad \mathsf{C}(\tau_0, \zeta_0) = -(\beta_0+1)\log(\beta_0\gamma_0) + \log\Gamma(\beta_0+1) + \frac{1}{2}\log|\Psi_0| \tag{1.55}$$

Plugging this into (1.26) we get that the marginal distribution of $\boldsymbol{x} = (x_1, \ldots, x_k)$ is given by

$$g_{NIG,k}(\boldsymbol{x}) = \frac{(\beta_0\gamma_0)^{\alpha_0}|\Psi_k|^{1/2}\Gamma(\alpha_k)}{(2\pi)^{dk/2}|\Psi_0|^{1/2}|\Sigma_0|^{k/2}\Gamma(\alpha_0)} \cdot \beta(\boldsymbol{x})^{-\alpha_k} \tag{1.56}$$

where $\Psi_k$ is defined by (1.35),

$$\alpha_k = \beta_0 + 1 + kd/2 \quad \text{and} \tag{1.57}$$

$$\beta(\boldsymbol{x}) = \beta_0\gamma_0 + \frac{1}{2}\sum_{i=1}^{k}(x_i - \overline{\boldsymbol{x}})^\top\Sigma_0^{-1}(x_i - \overline{\boldsymbol{x}}) + \frac{1}{2}(\overline{\boldsymbol{x}} - \mu_0)^\top k\Xi_k(\overline{\boldsymbol{x}} - \mu_0) \tag{1.58}$$

where

$$\Xi_k = (\Sigma_0 + k\Psi_0)^{-1} = \Psi_0^{-1}\Psi_k\Sigma_0^{-1}. \tag{1.59}$$

(the second equality in (1.59) is easily established by investigating its inverse). The detailed derivation of (1.56) from (1.26) can be found in Appendix A.1.3.

**Regularity.** Following the lines of the proof of regularity for the Normal-Inverse-Wishart model, in this case the proof boils down to establishing that if $\Sigma_0$ is a positive-definite matrix then the condition

$$\int_{\mathcal{X}} \exp\left\{-\frac{1}{2\lambda}(x-\mu)^\top \Sigma_0^{-1}(x-\mu)\right\} \mathrm{d}x < \infty \tag{1.60}$$

implies that $\lambda > 0$, which is straightforward.

**Convexity assumption.** By Theorem 1.2 the function $\boldsymbol{a}^\top \mathbf{B}\big(\theta(\eta)\big)$ is convex for any choice of $d \in \mathbb{N}$ and positive definite matrix $\Sigma_0$. For any fixed $\beta_0 > 0$ and positive definite matrix $\Psi$ we can deduce the convexity of ${\zeta_0}^\top \mathbf{B}\big(\theta(\eta)\big)$ by considering $d = 2$ and $\Sigma_0 = (\beta_0 + 2)\Psi_0$.

As a final point we note that Normal-Inverse-Gamma prior is a generalisation of the Normal prior in the sense that (1.50) becomes (1.28) for $\gamma_0 = 1$ and $\beta_0 \to \infty$.

# Chapter 2

# Geometry of MAP clustering and induced partitions

In this chapter we are concerned about two issues – the geometric properties of the MAP clustering (Section 2.1) and the approximation to the posterior score (1.8) when the data clustering is defined by a partition of the observation space and the data itself is and independently identically distributed sample with a given input distribution. This is explained in detail in Section 2.2.

In Section 2.1 we present the first important result of this dissertation, about 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 2.3 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.

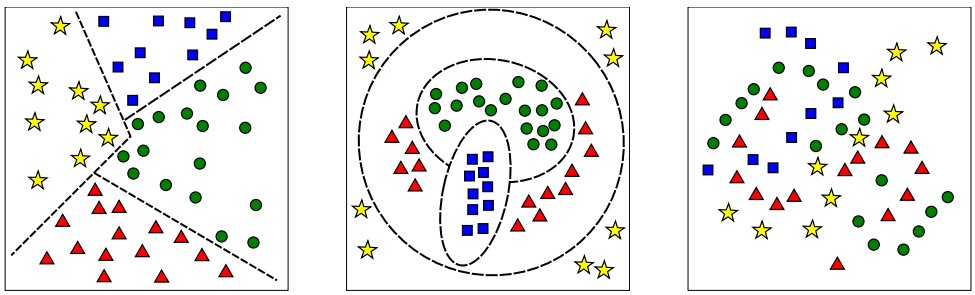
This separation result for the clusters of the MAP partition implies, loosely speaking, that the MAP clusters are contained within some decent 'chunks' of the observation space. This motivates us to 'reverse the optics' and consider clusterings (that we call *induced*) of the data that are defined by an a'priori fixed partition $\mathcal{A}$ of the observation space. We derive an 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 $\pi$ on the component probabilities (cf. (1.3)), provided that $\pi$ has a full support on the infinitely dimensional simplex $\triangle^{\infty}$.

## 2.1 Geometry of the MAP clustering

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

**Definition 2.1.** 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 2.2.** 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. Similarly, if $T(x) = [\mathrm{diag}(xx^\top), \mathrm{low}(xx^\top), x]$ and $\mathcal{Z}$ is $T$-linearly separable then every pair of elements of $\mathcal{Z}$ is separated (in geometric sense) by a quadratic surface. Hence, in this case we also use the term *quadratic separability.*



(a) This family is linearly separable.

(b) This family is quadratically separable. It is not linearly separable.

(c) This family is not quadratically separable.

Figure 2.1: 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).

**Notation.** For the notational convenience we will use the separability notions also with respect to the pairs of subsets or sequences of $\mathbb{R}^d$. For example, if $x_1, \ldots, x_n \in \mathbb{R}^d$ and $I, J$ are disjoint subsets of $[n]$ then the expression $\boldsymbol{x}_I$ *is linearly separated from* $\boldsymbol{x}_J$ means that the family $\big\{\{x_i \colon i \in I\}, \{x_j \colon j \in J\}\big\}$ is linearly separated.

**Theorem 2.3.** *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, described in Section 1.4, where the hyperparameter is identifiable. Then the family $\{\boldsymbol{x}_I \colon I \in \hat{\mathcal{I}}\}$ is $T$-linearly separable.*

Theorem 2.3 is a consequence of Lemma 2.4 and Lemma 2.5, which we state and prove below. A somewhat imprecise description of Lemma 2.4 is the following: it states that in order to prove the $\mathcal{L}$-separability property of the MAP clustering it is sufficient to show that for any two clusters of known sizes, if we allow a 'data exchange' (that preserves

the sizes) between these two clusters, the sum of inputs of the resulting clusters to the log-posterior score is maximised if they are $\mathcal{L}$-separated. Lemma 2.5 shows that linear separability occurs naturally in some general convex maximisation problem.

**Lemma 2.4.** *Let $\mathcal{L}$ be a family of real functions on $\mathbb{R}^d$. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $\hat{\mathcal{I}}$ be the MAP partition for $x_1, \ldots, x_n$ in some Bayesian Mixture Model, given by (1.6). For $\mathcal{U} \subset [n]$ and $0 < k < |\mathcal{U}|$ let $\hat{I}_{k,\mathcal{U}}$ be any subset of $\mathcal{U}$ of size $k$ that maximises $\log g_{\vartheta,k}(\boldsymbol{x}_I) + \log g_{\vartheta,l}(\boldsymbol{x}_{\mathcal{U}\setminus I})$, where $l = |\mathcal{U}| - k$ and $g_{\vartheta,k}$ is given by (1.5). In other words let*

$$\hat{I}_{k,\mathcal{U}} \in \operatorname*{argmax}_{I \subset \mathcal{U} : |I| = k} \big( \log g_{\vartheta,k}(\boldsymbol{x}_I) + \log g_{\vartheta,l}(\boldsymbol{x}_{\mathcal{U}\setminus I}) \big). \tag{2.1}$$

*If for $\mathcal{U} \subset [n]$, $0 < k < |\mathcal{U}|$ and any choice of $\hat{I}_{k,\mathcal{U}}$ the sets of observations $\boldsymbol{x}_{\hat{I}_{k,\mathcal{U}}}$ and $\boldsymbol{x}_{\mathcal{U}\setminus\hat{I}_{k,\mathcal{U}}}$ are separated by $\mathcal{L}$ then the whole family $\{\boldsymbol{x}_I : I \in \hat{\mathcal{I}}\}$ is separated by $\mathcal{L}$ (where $\hat{I}$ is MAP partition of $[n]$).*

*Proof.* Firstly note that by the definition of the MAP partition, using (1.8) and (1.7)

$$\hat{\mathcal{I}} \in \operatorname*{argmax}_{\text{partitions } \mathcal{I} \text{ of } [n]} \Big( \log \mathcal{P}_{\pi,n}(\mathcal{I}) + \sum_{I \in \mathcal{I}} \log g_{\vartheta,|I|}(\boldsymbol{x}_I) \Big). \tag{2.2}$$

Suppose that the assumptions of Lemma 2.4 hold. Suppose that $\hat{\mathcal{I}}$ is not separated by $\mathcal{L}$. Then there exist $\hat{I}, \hat{J} \in \hat{\mathcal{I}}$ such that $\boldsymbol{x}_{\hat{I}}$ and $\boldsymbol{x}_{\hat{J}}$ are not separated by $\mathcal{L}$. Let $\mathcal{U} = \hat{I} \cup \hat{J}$ and $k = |\hat{I}|$. Let $\tilde{I} = \hat{I}_{k,\mathcal{U}}$ and $\tilde{J} = \mathcal{U} \setminus \tilde{I}$. Moreover let $\tilde{\mathcal{I}}$ be a partition of $[n]$ obtained by replacing $\hat{I}, \hat{J}$ by $\tilde{I}, \tilde{J}$, i.e. $\tilde{\mathcal{I}} = \hat{\mathcal{I}} \setminus \{\hat{I}, \hat{J}\} \cup \{\tilde{I}, \tilde{J}\}$. Note that $\mathcal{P}_{\pi,n}(\hat{\mathcal{I}}) = \mathcal{P}_{\pi,n}(\tilde{\mathcal{I}})$ (we have $|\hat{I}| = |\tilde{I}|$ and $|\hat{J}| = |\tilde{J}|$, so we use the exchangeability of $\mathcal{P}_{\pi,n}$). Moreover $\boldsymbol{x}_{\hat{I}}$ and $\boldsymbol{x}_{\hat{J}}$ are not separated by $\mathcal{L}$ so by the assumptions of Lemma 2.4

$$\hat{I} \notin \operatorname*{argmax}_{I \subset \mathcal{U} : |I| = k} \big( \log g_{\vartheta,k}(\boldsymbol{x}_I) + \log g_{\vartheta,l}(\boldsymbol{x}_{\mathcal{U}\setminus I}) \big) \tag{2.3}$$

and hence, by the definition of $\tilde{I}$

$$\log g_{\vartheta,k}(\boldsymbol{x}_{\tilde{I}}) + \log f_l(\boldsymbol{x}_{\tilde{J}}) > \log g_{\vartheta,k}(\boldsymbol{x}_{\hat{I}}) + \log f_l(\boldsymbol{x}_{\hat{J}}). \tag{2.4}$$

This means that

$$\log \mathcal{P}_{\pi,n}(\tilde{\mathcal{I}}) + \sum_{I \in \tilde{\mathcal{I}}} \log g_{\vartheta,|I|}(\boldsymbol{x}_I) > \log \mathcal{P}_{\pi,n}(\hat{\mathcal{I}}) + \sum_{I \in \hat{\mathcal{I}}} \log g_{\vartheta,|I|}(\boldsymbol{x}_I), \tag{2.5}$$

which contradicts the definition of $\hat{\mathcal{I}}$ and the proof follows. □

**Lemma 2.5.** *Let $V \subseteq \mathbb{R}^D$ be a convex set. Let $f : V \to \mathbb{R}$ be a strictly concave function and $z_1, \ldots, z_{k+l} \in \mathbb{R}^D$ be pairwise distinct. If $\sum_{i \in I} z_i \in V$ for every $I \subseteq [k+l]$ such that $|I| = k$ and*

$$\hat{I} \in \operatorname*{argmin}_{I \subset [k+l] : |I| = k} f\big( \sum_{i \in I} z_i \big) \tag{2.6}$$

*then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{[k+l]\setminus\hat{I}}$ are linearly separable.*

*Proof.* Consider the set of all possible sums of $k$ distinct vectors $z_i$, i.e. $\mathcal{S}_k = \{\sum_{i\in I} z_i \colon I \subset [k+l], |I| = k\}$ and let $\hat{s}_k \in \operatorname{argmin}_{s\in\mathcal{S}_k} f(s)$. Since $f$ is strictly concave, then $\hat{s}_k$ is a vertex of the convex hull of $\mathcal{S}_k$, denoted by $\operatorname{conv}\mathcal{S}_k$ (see Figure 2.2). This means that there exist a vector $v_0 \in \mathbb{R}^D$ such that $\hat{s}_k$ is the furthest sum in the direction of $v_0$, or formally $\{\hat{s}_k\} = \operatorname{argmax}_{s\in\mathcal{S}_k}\langle s, v_0\rangle$ (cf. Moszyńska, 2005, Corollary 3.3.6), where $\langle\cdot,\cdot\rangle$ is the standard Euclidean scalar product. As the set of such vectors $v_0$ has a non-empty interior, we can also choose $v_0$ so that $\langle z_i, v_0\rangle$ are all different. Let $z_{(1)}, \ldots, z_{(k+l)}$ be an ordering of vectors $z_i$, decreasing 'in the direction $v_0$', i.e. $\{z_{(1)}, \ldots, z_{(k+l)}\} = \{z_1, \ldots, z_{k+l}\}$ and $\langle z_{(i)}, v_0\rangle > \langle z_{(j)}, v_0\rangle$ if $i < j$. Note that

$$\Big\langle \sum_{i\in I} z_i, v_0 \Big\rangle = \sum_{i\in I}\langle z_i, v_0\rangle \tag{2.7}$$

and therefore $\hat{I} = \{z_{(1)}, \ldots, z_{(k)}\}$. Thus the sets $\{z_i \colon i \in \hat{I}\}$ and $\{z_i \colon i \notin \hat{I}\}$ are linearly separated by the hyperplane $\{u \in \mathbb{R}^D \colon \langle u, v_0\rangle = \langle z_{(k)} + z_{(k+1)}, v_0\rangle/2\}$. □

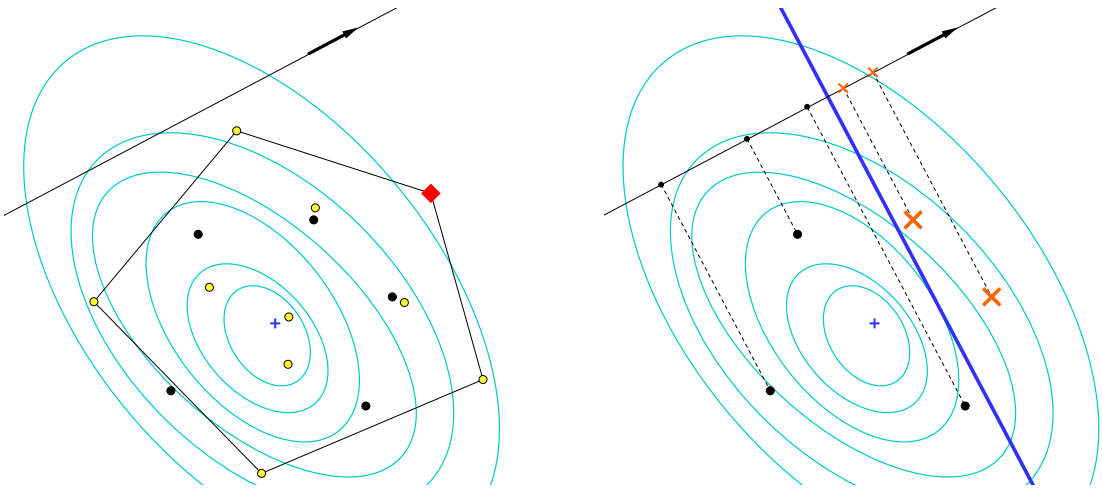


Figure 2.2: Illustration of the proof of Lemma 2.5. *Left panel:* The grey ellipses are the contour lines of the convex function $f$. In this example $n = 5$ and $k = 2$. Points $z_1, \ldots, z_5$ are marked as black points, and sums of their all possible pairs are marked as yellow dots and one red square, which is the minimiser of $f$ (the blue cross is the origin). The red square is a vertex of the convex hull of possible sums, and as such is a point furthest in some direction, also marked on the picture. *Right panel:* As the red square was the furthest in the highlighted direction, it is the sum of two points furthest in that direction (orange crosses) and therefore those two points are separated from others by a line perpendicular in that direction (marked in blue).

*Proof of Theorem 2.3.* Let $\mathcal{U} \subseteq [n]$, $k, l \in \mathbb{N}$ and $\hat{I}_{k,\mathcal{U}}$ be as in Lemma 2.4. Plugging the

formula (1.26) into (2.1) gives:

$$\begin{aligned}\hat{I}_{k,\mathcal{U}} \in \operatorname*{argmax}_{I\subset\mathcal{U}\colon |I|=k} \Big( \sum_{i\in I} \log h(x_i) + \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_I}, \boldsymbol{\zeta}_k) - \mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta}) + \\ + \sum_{i\in\mathcal{U}\setminus I} \log h(x_i) + \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_{\mathcal{U}\setminus I}}, \boldsymbol{\zeta}_l) - \mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta}) \Big) = \\ = \operatorname*{argmax}_{I\subset\mathcal{U}\colon |I|=k} \Big( \sum_{i\in\mathcal{U}} \log h(x_i) + \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_I}, \boldsymbol{\zeta}_k) + \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_{\mathcal{U}\setminus I}}, \boldsymbol{\zeta}_l) - 2\mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta}) \Big) = \\ = \operatorname*{argmax}_{I\subset\mathcal{U}\colon |I|=k} \Big( \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_I}, \boldsymbol{\zeta}_k) + \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_{\mathcal{U}\setminus I}}, \boldsymbol{\zeta}_l) \Big)\end{aligned} \tag{2.8}$$

Let $\boldsymbol{t}_i = \begin{bmatrix} T(x_i) \\ \boldsymbol{a} \end{bmatrix}$ and let $\boldsymbol{t}_0 = \begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{\zeta} \end{bmatrix}$ and $\boldsymbol{t}_{\mathcal{U}} = \sum_{i\in\mathcal{U}} \boldsymbol{t}_i$. By the assumed identifiability of the hyperparameter and Bickel and Doksum (2015, Theorem 1.6.4), $\mathsf{C}$ is a strictly convex function. Hence the functions $\underline{f}(\boldsymbol{t}) = \mathsf{C}(\boldsymbol{t}_0 + \boldsymbol{t})$ and $\overline{f}(\boldsymbol{t}) = \mathsf{C}(\boldsymbol{t}_0 + \boldsymbol{t}_{\mathcal{U}} - \boldsymbol{t})$ are also strictly convex and so is their sum, $f(\boldsymbol{t}) = \underline{f}(\boldsymbol{t}) + \overline{f}(\boldsymbol{t})$. Note that

$$\underline{f}\Big(\sum_{i\in I} \boldsymbol{t}_i\Big) = \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_I}, \boldsymbol{\zeta}_k) \quad \text{and} \quad \overline{f}\Big(\sum_{i\in I} \boldsymbol{t}_i\Big) = \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_{\mathcal{U}\setminus I}}, \boldsymbol{\zeta}_l) \tag{2.9}$$

and hence by (2.8) we get

$$\hat{I}_{k,\mathcal{U}} \in \operatorname*{argmax}_{I\subset\mathcal{U}\colon |I|=k} f\Big(\sum_{i\in I} \boldsymbol{t}_i\Big). \tag{2.10}$$

Therefore by Lemma 2.5 we obtain that $\boldsymbol{t}_{\hat{I}_{k,\mathcal{U}}}$ and $\mathbf{t}_{\mathcal{U}\setminus\hat{I}_{k,\mathcal{U}}}$ are linearly separable. This yields $T$-linear separability of $\boldsymbol{x}_{\hat{I}_{k,\mathcal{U}}}$ and $\boldsymbol{x}_{\mathcal{U}\setminus\hat{I}_{k,\mathcal{U}}}$ and the proof follows. □

We now list three Corollaries that follow from Theorem 2.3 and the formula for the sufficient statistic in the Normal models, described in Section 1.4.1. The fact that the hyperparameter is identifiable in these models is straightforward.

**Corollary 2.6.** *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 Normal-Normal Bayesian Mixture Model. Then the family $\{\boldsymbol{x}_I \colon I \in \hat{\mathcal{I}}\}$ is linearly separable.*

**Corollary 2.7.** *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 Normal-Inverse-Wishart Bayesian Mixture Model. Then the family $\{\boldsymbol{x}_I \colon I \in \hat{\mathcal{I}}\}$ is quadratically separable.*

**Corollary 2.8.** *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 Normal-Inverse-Gamma Bayesian Mixture Model. Then the family $\{\boldsymbol{x}_I \colon I \in \hat{\mathcal{I}}\}$ is quadratically separable. Moreover, in this case every two clusters are separated (in standard, geometric sense) by a multidimensional ellipse.*

### 2.1.1 Analogy to the properties of the Fisher Discriminant Analysis

Here we would like to draw attention to the aesthetic analogy of Theorem 2.3 to the properties of the classical Fisher's Linear or Quadratic Discriminant Analysis.

Suppose we have access to samples from two heterogenous populations: $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_l$. Then $z$ is observed and we need to classify it as coming from the population of $x$'es or $y$'s. Fisher Discriminant Analysis approach is to assume that both initial samples are taken from some multivariate normal distributions. If the population parameters were known and equal to $(\mu_1, \Sigma_1)$ and $(\mu_2, \Sigma_2)$ respectively, and the population frequecies were $p_1$ and $p_2$ then it is a natural (and optimal in the Bayesian sense, cf. Bishop, 2006, Section 1.5.1) decision to classify the new observation to the first group if and only if $p_1 g_{(\mu_1,\Sigma_1)}(z) > p_2 g_{(\mu_2,\Sigma_2)}(z)$, where $g_{(\mu,\Sigma)}$ is the density of the $\mathcal{N}(\mu, \Sigma)$ distribution. This easily translates to the following inequality:

$$\frac{1}{2}(z-\mu_2)^\top \Sigma_2^{-1}(z-\mu_2) - \frac{1}{2}(z-\mu_1)^\top \Sigma_1^{-1}(z-\mu_1) > \log\frac{p_2}{p_1} - \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}. \tag{2.11}$$

Since we do not have the access to the true theoretical values of $p_i$, $\mu_i$, $\Sigma_i$ $(i = 1, 2)$, we replace them by their empirical estimates $\hat{\mu}_i$ and $\hat{\Sigma}_i$. This does not change the fact that left-hand side of (2.11) is a quadratic function in $z$ and hence the boundaries of the decision region are also quadratic surfaces.

If for some reasons we expect the two populations to have the same theoretical covariance $\Sigma = \Sigma_1 = \Sigma_2$, instead of two separate estimators $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ we can consider one pooled estimator $\hat{\Sigma}$. Then (2.11) simplifies to

$$(\hat{\mu}_1 - \hat{\mu}_2)^\top \hat{\Sigma}^{-1} z > \log\frac{\hat{p}_2}{\hat{p}_1}, \tag{2.12}$$

the left-hand side being a linear function of $z$ and hence the decision boundaries are now simply hyperplanes. Thus we arrived at conclusion similar to that implied by our results concerning the properties of the MAP clustering: with assumption of normality, when the population covariances are assumed to be the same, the boundaries are linear surfaces, otherwise they are quadratic.

This analogy can be easily extended to the case of general exponential families. If we assume that the density $g_\theta$ is given by (1.16), the comparison $p_1 g_{\hat{\theta}_1}(z) > p_2 g_{\hat{\theta}_1}(z)$ can be expressed as

$$\big(\eta(\hat{\theta}_1) - \eta(\hat{\theta}_2)\big)^\top T(z) > \log\frac{p_2}{p_1} + \mathsf{B}(\hat{\theta}_1) - \mathsf{B}(\hat{\theta}_2) \tag{2.13}$$

which implies that the decision boundaries are the contour lines of a linear functional of the sufficient statistics $T$.

It should, however, be stressed that Fisher Discriminant Analysis concerns a different domain of statistical questions than those considered in this dissertation, namely *classification*, not cluster analysis. Moreover, the tools used to prove Theorem 2.3 were slightly more involved; among others, it used the convexity of the log-partition function $\mathsf{C}$, whereas the separability observation for the discriminant analysis is a straightforward consequence of the formula (1.16).

## 2.2 Induced partitions

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' (1.8). 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 normals (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 2.9.** 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 disjoint 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 2.10.** 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 $n \in \mathbb{N}$.

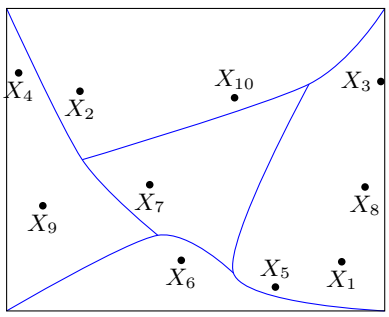



Figure 2.3: In this picture the observation space $\mathcal{X}$ is the rectangle and the partition $\mathcal{A}$ is defined by the blue separation curves. The points $X_1, \ldots, X_{10}$ are drawn uniformly from $\mathcal{X}$. The random partition of $\{1, 2, \ldots, 10\}$ induced by $\mathcal{A}$ is

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

According to Remark 2.10, 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 (1.8). In order to specify the limit, we recall the notion of *convex conjugate*.

**Definition 2.11.** 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 2.12.** *Consider the infinite conjugate exponential Bayesian Mixture Model, in which the component measures are given by* (1.16) *and the base measure is given by* (1.21)*. Suppose that the exponential family is regular and that the convexity assumption* (1.27) *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* $\mathsf{A}^*$ *is the convex conjugate of the log-partition function* $\mathsf{A}$*, given by* (1.19)*,*

*(ii)* $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big) \in \operatorname{int}\Omega$ *for some $r \in \mathbb{N}$, where $\Omega$ is the natural hyperparameter space, defined by* (1.23)*.*

*Let $Q$ be the joint probability function given by* (1.8)*, in which $g_{\vartheta,k}$ is given by* (1.26)*. Let $X_1, X_2, \ldots \overset{iid}{\sim} P$ and $\mathbf{X}_{1:n} = (X_1, \ldots, X_n)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}}(\mathbf{X}_{1:n})\big) \overset{a.s.}{=} E_P \log h(X) + \Delta_P(\mathcal{A}) \tag{2.14}$$

*where*

$$\Delta_P(\mathcal{A}) = \sum_{A\in\mathcal{A}} P(A) \cdot \mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) + \sum_{A\in\mathcal{A}} P(A) \log P(A). \tag{2.15}$$

The remaining part of this section is devoted to the proof of Theorem 2.12. But now let us point out its obvious consequence.

**Corollary 2.13.** *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 2.12 almost surely there exists $N$ such that*

$$Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}_1}(\mathbf{X}_{1:n})\big) > Q\big(\mathbf{X}_{1:n}, \mathcal{J}_n^{\mathcal{A}_1}(\mathbf{X}_{1:n})\big) \quad \text{for } n > N \tag{2.16}$$

Hence, as long as the induced partitions are concerned, the $\Delta_P$ function is an indicator of which of these partitions gives larger posterior score given by (1.8), when our data is an independent sample from the probability distribution $P$. In this sense we can hope that $\Delta_P$ 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. Section 2.1), 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 Chapter 3). For now, let us move to the proof of Theorem 2.12.

### 2.2.1 Proof of Theorem 2.12

We start with a well known fact from the Functional Analysis. Since its proof is nice and short, we present it here for completeness. Let us recall that if $\mu$ is a measure on some space $\mathcal{X}$ and $p \geq 1$ then by $L^p(\mathcal{X}, \mu)$ we denote the space of all real functions on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} |f(x)|^p \mathrm{d}\mu(x) < \infty$, equipped with the $p$-norm

$$\|f\|_p = \left( \int_{\mathcal{X}} |f(x)|^p \mathrm{d}\mu(x) \right)^{\frac{1}{p}}. \tag{2.17}$$

Moreover, $L^\infty(\mathcal{X}, \mu)$ is the space of functions with finite *essential supremum* (a quantity that at the same time is the norm in this space), defined by

$$\|f\|_\infty = \inf\{C \geq 0 \colon |f(x)| \leq C \text{ for } \mu\text{-almost every } x\}. \tag{2.18}$$

**Lemma 2.14.** *Let $(\mathcal{X}, \mu)$ be a measurable space. If $f \in L^{n_0}(\mathcal{X}, \mu) \cap L^\infty(\mathcal{X}, \mu)$ for some $n_0 > 0$ then $\lim_{n\to\infty} \|f\|_n = \|f\|_\infty$.*

*Proof.* Let $\varepsilon > 0$ and let $M = \|f\|_\infty$ and $A_\varepsilon = \{x \colon f(x) > M - \varepsilon\}$. Clearly $0 < \mu(A_\varepsilon) < \infty$. Moreover

$$\|f\|_n^n \geq \int_{A_\varepsilon} f^n \mathrm{d}\mu \geq \mu(A_\varepsilon)(M - \varepsilon)^n. \tag{2.19}$$

Hence $\liminf_{n\to\infty} \|f\|_n \geq M - \varepsilon$ and, by the arbitrary choice of $\varepsilon$, $\liminf_{n\to\infty} \|f\|_n \geq M$.

On the other hand, for $n > n_0$ we have

$$\|f\|_n^n = \int_{A_1} f^n \mathrm{d}\mu + \int_{\mathcal{X}\setminus A_1} f^n \mathrm{d}\mu \leq M^n \mu(A_1) + \|f\|_{n_0}^{n_0}, \tag{2.20}$$

which gives $\limsup_{n\to\infty} \|f\|_n \leq M$ and the proof follows. □

Lemma 2.14 plays central role in the approximation of the $n$-th root of both factors in (1.8). However, in both cases the situation is slightly more involved that the one presented by Lemma 2.14, since in fact we will deal with a sequence of functions $f_n$ converging pointwise to $f$ and our goal will be to show that $\|f_n\|_n \to \|f\|_\infty$. The proof of this convergence requires some additional steps and observations.

We split the analysis of the asymptotic limit of the logarithm of the function $Q$ given by (1.8) into two parts: the asymptotic limit of the logarithm of $\mathcal{P}_{\pi,n}$ and the asymptotic limit of the logarithm of $g_{\vartheta,n}$.

**Asymptotic limit of** $\log \mathcal{P}_{\pi,n}$ This aspect is summarised by Lemma 2.15.

**Lemma 2.15.** *Let $P$ be a probability distribution on $\mathbb{R}^d$ and let $\mathcal{A}$ be a* finite *$P$-partition of the observation space. Let $X, X_1, X_2, \ldots \overset{iid}{\sim} P$ and $\mathbf{X}_{1:n} = (X_1, \ldots, X_n)$ Let $\mathcal{P}_{\pi,n}$ be a probability distribution on the partitions of $[n]$, generated by the probability distribution $\pi$ on $\triangle^\infty$ with a full support. Then*

$$\lim_{n\to\infty} \sqrt[n]{\mathcal{P}_{\pi,n}\big(\mathcal{I}_n^{\mathcal{A}}(\mathbf{X}_{1:n})\big)} \overset{a.s.}{=} \prod_{A\in\mathcal{A}} P(A)^{P(A)} \tag{2.21}$$

Lemma 2.15 follows simply from Lemma 2.16 and the Strong Law of Large Numbers.

**Lemma 2.16.** *Let $\mathcal{P}_{\pi,n}$ be a probability distribution on the partitions of $[n]$, generated by the probability distribution $\pi$ on $\triangle^\infty$ with a full support. Fix $K \in \mathbb{N}$ and consider a sequence of partitions $(\mathcal{I}_n)_{n\in\mathbb{N}}$, where $\mathcal{I}_n = \{I_{n,1}, \ldots, I_{n,K}\}$ is a partition of $[n]$ (it is possible that $I_{n,i} = \emptyset$ for some $i \le K$). Assume that $|I_{n,k}|/n \to \alpha_k > 0$ for $k \le K$. Then*

$$\lim_{n\to\infty} \sqrt[n]{\mathcal{P}_{\pi,n}(\mathcal{I}_n)} = \prod_{k=1}^{K} \alpha_k^{\alpha_k} \tag{2.22}$$

In order to prove Lemma 2.16, we need a closed-form expression for $\mathcal{P}_{\pi,n}$. This is given by Lemma 2.17

**Lemma 2.17.** *Let $\pi$ be a probability distribution on $\triangle^\infty$ that generates the probability $\mathcal{P}_{\pi,n}$ on the space of partitions of $[n]$. Then for every partition $\mathcal{I}$ of $[n]$*

$$\mathcal{P}_{\pi,n}(\mathcal{I}) = \int_{\triangle^\infty} \sum_{\psi\colon \mathcal{I} \overset{1-1}{\to} \mathbb{N}} \prod_{I\in\mathcal{I}} p_{\psi(I)}^{|I|} \mathrm{d}\pi(\boldsymbol{p}) \tag{2.23}$$

*where sum ranges over all injective functions from $\mathcal{I}$ to $\mathbb{N}$.*

*Proof.* The formula (2.23) may seem rather complicated at first, but its justification is straightforward. Let us go back to (1.4) and suppose that the weights $\boldsymbol{p} = (p_i)_{i=1}^\infty$ and the atoms $\boldsymbol{\theta} = (\theta_i)_{i=1}^\infty$ are fixed. We need to know what is the probability that $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \,|\, \boldsymbol{p}, \boldsymbol{\theta} \overset{\text{iid}}{\sim} \sum_{i=1}^m p_i \delta_{\theta_i}$ induces a partition $\mathcal{I}$. This would mean that for every $I \in \mathcal{I}$ all the values $\phi_i$ for $i \in I$ are equal to $\theta_j$ for some $j \in \mathbb{N}$; let $j = \psi(I)$. The values $\psi(I)$ must be different for different $I \in \mathcal{I}$, otherwise $\mathcal{I}$ would not be generated. The probability of the sequence $(\phi_1, \ldots, \phi_n)$ where $\phi_i = \theta_{\psi(I)}$ for $i \in I$ is equal to $\prod_{I\in\mathcal{I}} p_{\psi(I)}^{|I|}$. Since any assignment of clusters to atoms is valid, so for fixed $\boldsymbol{p}$ the probability of $\mathcal{I}$ is equal to $\sum_{\psi\colon \mathcal{I} \overset{1-1}{\to} \mathbb{N}} \prod_{I\in\mathcal{I}} p_{\psi(I)}^{|I|}$. Since $\boldsymbol{p} \sim \pi$ is random, we have to integrate it out and (2.23) follows. □

*Proof of Lemma 2.16.* Firstly note that for sufficiently large $n$ we have $|I_{k,n}| \ge 1$ for all $k \le K$. Then in (2.23) we sum functions that depend on exactly $K$ coordinates of $\boldsymbol{p}$. Let

$$\blacktriangle^K = \{(p_1, \ldots, p_K)\colon \sum_{i=1}^{K} p_i \le 1, \forall_i p_i \ge 0\}. \tag{2.24}$$

For any $\psi\colon [K] \overset{1-1}{\to} \mathbb{N}$ let $\pi_\psi$ be a probability measure on $\blacktriangle^K$ defined by

$$\pi_\psi(A) = \pi\Big(\big\{(p_1, p_2, \ldots) \in \triangle^\infty\colon (p_{\psi(1)}, p_{\psi(2)}, \ldots, p_{\psi(K)}) \in A\big\}\Big). \tag{2.25}$$

In other words, $\pi_\psi$ is a push-forward of the $\pi$ measure in the projection on the coordinates $(p_{\psi(1)}, p_{\psi(2)}, \ldots, p_{\psi(K)})$. Now let $\sigma$ be a measure on $\blacktriangle^K$ defined by

$$\sigma = \sum_{\psi\colon [K] \overset{1-1}{\to} \mathbb{N}} \pi_\psi. \tag{2.26}$$

Note that since every summand on the right-hand side of (2.26) is a probability measure, the measure $\sigma$ is not a finite measure on $\blacktriangle^K$.

Now we can express (2.23) in the form of an integral on the $K$-dimensional set $\blacktriangle^K$ as

$$\mathcal{P}_{\pi,n}(\mathcal{I}_n) = \int_{\blacktriangle^K} \prod_{k=1}^{K} p_k^{|I_{k,n}|} \mathrm{d}\sigma(p_1, \ldots, p_K). \tag{2.27}$$

Hence

$$\sqrt[n]{\mathcal{P}_{\pi,n}(\mathcal{I}_n)} = \sqrt[n]{\int_{\blacktriangle^K} \prod_{k=1}^{K} p_i^{|I_{k,n}|} \mathrm{d}\sigma(p_1, \ldots, p_K)} = \|g_n\|_n \tag{2.28}$$

where $g_n(p_1, \ldots, p_K) = \prod_{k=1}^{K} p_k^{|I_{k,n}|/n}$ and $\|\cdot\|_n$ is the norm in $L^n(\blacktriangle^K, \sigma)$ space.

Since $\sigma$ is not a finite measure on $\blacktriangle^K$, in the remaining part of the proof we will have to be careful that the functions we are considering belong to the space $L^n(\blacktriangle^K, \sigma)$ for sufficiently large $n$.

Let $g(p_1, \ldots, p_K) = \prod_{k=1}^{K} p_k^{\alpha_k}$ and let $h(p_1, \ldots, p_K) = \prod_{k=1}^{K} p_k$. Note that

$$\int_{\blacktriangle^K} h(p_1, \ldots, p_K) \mathrm{d}\sigma(p_1, \ldots, p_K) = \mathcal{P}_{\pi,K}\Big(\{\{1\}, \{2\}, \ldots, \{K\}\}\Big) \leq 1. \tag{2.29}$$

Moreover for $n > 1/\min \alpha_i$ we have $g^n(p_1, \ldots, p_K) \leq h(p_1, \ldots, p_K)$ and therefore $g \in L^n(\blacktriangle^K, \sigma)$ for $n > 1/\min \alpha_i$. Because $g$ is bounded by 1, we can use Lemma 2.14 to obtain

$$\lim_{n\to\infty} \|g\|_n = \|g\|_\infty = \sup_{\blacktriangle^K} g = \prod_{k\leq K} \alpha_k^{\alpha_k}. \tag{2.30}$$

The equality $\|g\|_\infty = \sup_{\blacktriangle^K} g$ is a consequence of the assumption that $\pi$ has a full support on $\triangle^\infty$. The equality $\sup_{\blacktriangle^K} g = \prod_{k\leq K} \alpha_k^{\alpha_k}$ follows in a standard way from the Lagrange multipliers, the details are left for Lemma A.2 from the Appendix.

We now prove that $\|g_n - g\|_n \to 0$. It is not a straightforward consequence of the pointwise convergence of $g_n$ to $g$ since $\sigma$ is not a finite measure on $\blacktriangle^K$.

Clearly, $(|I_{k,n}|/n - \alpha_k/2) \to \alpha_k/2 > 0$ and hence $g_n g^{-1/2} \to g^{1/2}$ pointwise on $\blacktriangle^K$. As $\blacktriangle^K$ is compact, we have uniform convergence as well, which means $\|g_n g^{-1/2} - g^{1/2}\|_\infty \to 0$ on $\blacktriangle^K$.

Let $N \in \mathbb{N}$ be chosen so that for $n > N$ we have $\|g_n g^{-1/2} - g^{1/2}\|_\infty < \varepsilon$ and $n\alpha_k \geq 2$ for $k \leq K$. Then for $n > N$ we have $g^{n/2} \leq h$ and

$$\begin{aligned} \|g_n - g\|_n^n = \int_{\blacktriangle^K} |g_n - g|^n \mathrm{d}\sigma = \int_{\blacktriangle^K} |g_n g^{-1/2} - g^{1/2}|^n g^{n/2} \mathrm{d}\sigma \leq \\ \leq \epsilon^n \int_{\blacktriangle^K} g^{n/2} \mathrm{d}\sigma \leq \epsilon^n \int_{\blacktriangle^K} h \mathrm{d}\sigma \leq \epsilon^n, \end{aligned} \tag{2.31}$$

hence

$$\lim_{n\to\infty} \|g_n - g\|_n = 0. \tag{2.32}$$

The result follows from (2.30), (2.32) and the triangle inequality:

$$\big|\|g_n\|_n - \|g\|_\infty\big| \leq \big|\|g_n\|_n - \|g\|_n\big| + \big|\|g\|_n - \|g\|_\infty\big| \leq \|g_n - g\|_n + \big|\|g\|_n - \|g\|_\infty\big|. \tag{2.33}$$

□

**Asymptotic limit of** $\log g_{\vartheta,n}$ The goal of this paragraph is to prove Lemma 2.18, which together with Lemma 2.15 will easily imply Theorem 2.12.

**Lemma 2.18.** *If $P$ is a probability distribution on $\mathbb{R}^d$, $\mathcal{A}$ is a finite $P$-partition of $\mathbb{R}^d$ and $X_1, X_2, \ldots \overset{iid}{\sim} P$. Let $g_{\vartheta,k}$ be the marginal density defined by (1.26) in the regular exponential family and suppose that the convexity assumption (1.27) holds. Suppose that $E_P \log h(X) < \infty$ and for every $A \in \mathcal{A}$*

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

*(ii)* $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big) \in \operatorname{int}\Omega$ *for some* $r \in \mathbb{N}$.

*Then*

$$\lim_{n\to\infty} \frac{1}{n} \log g_{\vartheta,n}\big(X_{1:n} \,|\, \mathcal{I}_n^{\mathcal{A}}(X_{1:n})\big) \overset{a.s.}{=} E_P \log h(X) + \sum_{A\in\mathcal{A}} P(A) \cdot \mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) \tag{2.34}$$

The proof is an easy consequence of Proposition 2.19, proved later.

**Proposition 2.19.** *Consider the regular exponential family, described in Section 1.4, and suppose that it satisfies the convexity assumption (1.27). Let $x_1, x_2, \ldots$ be a sequence of points in* $\operatorname{supp} h$ *such that* $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n T(x_i) = t_0 \in \mathbb{R}^p$. *Suppose that* $\mathsf{A}^*(t_0) < \infty$ *and for some $r \in \mathbb{N}$ we have* $(rt_0, r\boldsymbol{a}) \in \operatorname{int}\Omega$. *Then*

$$\lim_{n\to\infty} \frac{1}{n} \mathsf{C}(\boldsymbol{\tau}_{\boldsymbol{x}_{1:n}}, \boldsymbol{\zeta}_n) = \mathsf{A}^*(t_0). \tag{2.35}$$

*Proof of Lemma 2.18 using Proposition 2.19.*

$$\begin{aligned}
\log g_{\vartheta,n}(X_{1:n} \,|\, \mathcal{I}_n^{\mathcal{A}}) &= \sum_{I\in\mathcal{I}_n^{\mathcal{A}}} \log g_{\vartheta,|I|}(X_I) = \\
&= \sum_{i=1}^n \log h(X_i) + \sum_{I\in\mathcal{I}_n^{\mathcal{A}}} \mathsf{C}(\boldsymbol{\tau}_{X_I}, \boldsymbol{\zeta}_{|I|}) - |\mathcal{A}| \cdot \mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta}) = \\
&= \sum_{i=1}^n \log h(X_i) + \sum_{A\in\mathcal{A}} \mathsf{C}(\boldsymbol{\tau}_{X_{I_n^A}}, \boldsymbol{\zeta}_{|I_n^A|}) - |\mathcal{A}| \cdot \mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta})
\end{aligned} \tag{2.36}$$

By the Strong Law of Large Numbers $\frac{|\mathcal{I}_n^A|}{n} \to P(A)$ almost surely and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log h(X_i) = E_P \log h(X) \tag{2.37}$$

Again, by the Strong Law of Large Numbers, almost surely for every $A \in \mathcal{A}$:

$$\lim_{n\to\infty} \frac{\sum_{i\in I_n^A} T(X_i)}{|I_n^A|} = E_P(T(X) \,|\, X \in A) \tag{2.38}$$

Hence by (2.36), (2.38) and Proposition 2.19

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log f(X_{1:n} \,|\, \mathcal{I}_n^{\mathcal{A}}) = \\
&= \lim_{n\to\infty} \Big( \frac{1}{n} \sum_{i=1}^{n} \log h(X_i) + \sum_{A\in\mathcal{A}} \frac{|I_n^A|}{n} \cdot \frac{1}{|I_n^A|} \mathsf{C}(\boldsymbol{\tau}_{X_{I_n^A}}, \boldsymbol{\zeta}_{|I_n^A|}) - \frac{|\mathcal{A}|}{n} \cdot \mathsf{C}(\boldsymbol{\tau}, \boldsymbol{\zeta}) \Big) = \\
&= E_P \log h(X) + \sum_{A\in\mathcal{A}} P(A) \cdot \mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big).
\end{aligned} \tag{2.39}$$

□

To prove Proposition 2.19 we also need to use Lemma 2.14, but again it is not a straightforward application and requires some additional steps.

*Proof of Proposition 2.19.* Let

$$\varphi_n(\theta) = \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} \tau_{\boldsymbol{x}_{1:n}}/n \\ \zeta_n/n \end{bmatrix} \right\} \tag{2.40}$$

Consider a measurable space $(\Theta, \Psi)$, where $\Psi$ is a measure on $\Theta$ with density $\psi(\theta)$ with respect to the Lebesgue measure. In this proof we will consider the spaces $\mathcal{L}^p(\Theta, \Psi)$, $p \geq 1$, with their norms $\|\cdot\|_p$. It is clear from the formula (1.22) that

$$\frac{1}{n} \mathsf{C}(\tau_{\boldsymbol{x}_{1:n}}, \zeta_n) = \log \|\varphi_n\|_n. \tag{2.41}$$

Now let

$$\varphi(\theta) = \exp\left\{ [\eta(\theta)^\top, -\mathbf{B}(\theta)^\top] \begin{bmatrix} t_0 \\ \boldsymbol{a} \end{bmatrix} \right\}. \tag{2.42}$$

Note that, using the regularity assumption

$$\|\varphi\|_\infty = \sup_{\theta\in\Theta} \exp\left\{ \eta(\theta)^\top t_0 - \mathsf{B}(\theta)^\top \right\} = \sup_{\eta\in\mathcal{E}} \exp\left\{ \eta^\top t_0 - \mathsf{A}(\eta) \right\} = \exp\{\mathsf{A}^*(t_0)\}. \tag{2.43}$$

where $\mathcal{E}$ is the natural parameter space, defined by (1.20). By (2.41) and (2.43) to conclude the proof it is enough to show that $\|\varphi_n\|_n \to \|\varphi\|_\infty$. According to Lemma A.3 (stated and proved in the Appendix), to ensure this convergence it is enough to check the following conditions:

- $\boxed{\varphi, \varphi_n \geq 0}$ This is straightforward from the definition.
- $\boxed{\|\varphi\|_\infty < \infty}$ This follows from (2.43) and the assumptions.
- $\boxed{\|\varphi\|_r < \infty \text{ and } \|\varphi_n\|_r \to \|\varphi\|_r}$ It is a consequence of

$$\|\varphi_n\|_r^r = \exp\Big\{ \mathsf{C}\big(\frac{r\tau_{\boldsymbol{x}_{1:n}}}{n}, \frac{r\zeta_n}{n}\big) \Big\}, \quad \|\varphi\|_r^r = \exp\Big\{ \mathsf{C}\big(rt_0, r\boldsymbol{a}\big) \Big\}. \tag{2.44}$$

  The function $\mathsf{C}$ is convex in $\Omega$ and hence continuous in $\operatorname{int}\Omega$ (Rockafellar, 1970, Theorem 10.1). As $(rt_0, r\boldsymbol{a}) \in \operatorname{int}\Omega$ and $\big(\frac{r\tau_{\boldsymbol{x}_{1:n}}}{n}, \frac{r\zeta_n}{n}\big) \to (rt_0, r\boldsymbol{a})$, we get $\|\varphi\|_r^r < \infty$ and $\|\varphi_n\|_r^r \to \|\varphi\|_r^r$.

- $\boxed{\|\varphi_n - \varphi\|_\infty \to 0}$ More verbosely, this is

$$\exp\left\{\frac{\tau_{\boldsymbol{x}_{1:n}}^\top}{n}\eta(\theta) - \frac{\zeta_n^\top}{n}\mathbf{B}(\theta)\right\} \overset{L^\infty(\Theta)}{\to} \exp\left\{t_0^\top\eta(\theta) - a^\top\mathbf{B}(\theta)\right\}, \tag{2.45}$$

which, by the regularity assumption, is equivalent to

$$\exp\left\{\frac{\tau_{\boldsymbol{x}_{1:n}}^\top}{n}\eta - \mathsf{A}_n(\eta)\right\} \overset{L^\infty(\mathcal{E})}{\to} \exp\left\{t_0^\top\eta - \mathsf{A}(\eta)\right\}. \tag{2.46}$$

in which $\mathsf{A}_n(\eta) = \frac{\zeta_n^\top}{n}\mathbf{B}(\theta(\eta)) = \frac{\zeta_0^\top}{n}\mathbf{B}(\theta(\eta)) + \mathsf{A}(\eta)$ is a convex function, by the convexity assumption (1.27) and the fact that a sum of convex functions is a convex function itself. Of course in (2.46) we have a pointwise convergence, but the space $\mathcal{E}$ can (and, in reasonable cases, is) unbounded and hence to establish uniform convergence we need to make some additional observations. They will base mostly on the convexity of the log-partition function $\mathsf{A}$.

Let $l(\eta) = t_0^\top\eta - \mathsf{A}(\eta)$; it is a concave function, since the log-partition function function $\mathsf{A}$ is convex. Let $l_n(\eta) = \frac{\tau_{\boldsymbol{x}_{1:n}}^\top}{n}\eta - \mathsf{A}_n(\eta)$; it is a concave function by the same reasoning. By Corollary 2.3.1 in Bickel and Doksum (2015) there exists $\hat{\eta} \in \mathcal{E}$ such that $l(\hat{\eta}) = \sup_{\eta\in\mathcal{E}} l(\eta) =: M$. Fix $\varepsilon > 0$. Let $F = \{\eta \in \mathcal{E} \colon l(\eta) \geq \log\varepsilon\}$. Without loss of generality we can assume that $\log\varepsilon < M$ so that $M \in F$. Since $\mathcal{E} = \{\eta \colon l(\eta) > -\infty\}$, $F$ is a compact subset of $\mathcal{E}$ (the details of this implication are left for Lemma A.5). Since $l_n$ converges to $l$ pointwise and $F$ is compact, by the Dini's theorem we have an uniform convergence of $l_n$ to $l$ on $F$. Hence there exist $N_1 \in \mathbb{N}$ such that $|l_n(\eta) - l(\eta)| < \varepsilon$ for $n > N_1$ and $\eta \in F$. In the same way we can prove that there exists $N_2 \in N$ such that $|e^{l_n(\eta)} - e^{l(\eta)}| < \varepsilon$ for $n > N_2$ and $\eta \in F$. Let $N = \max\{N_1, N_2\}$. For $\eta \in \mathcal{E} \setminus F$ we have $l(\eta) < \varepsilon$. Note that for $n > N$ we have $l_n(\eta) < l(\eta) + \varepsilon = \log\varepsilon + \varepsilon$ for $\eta \in \partial F$ and $l_n(\hat{\eta}) > l(\hat{\eta}) - \varepsilon = M - \varepsilon$. Again, we will not lose generality by assuming that $\log\varepsilon + 2\varepsilon < M$, so that $\log\varepsilon + \varepsilon < M - \varepsilon$. It then follows from the concavity of $l_n$ that $l_n(\eta) < \log\varepsilon + \varepsilon$ for $\eta \in \mathcal{E} \setminus F$. Therefore for $n > N$ we have

$$\begin{aligned} |e^{l_n(\eta)} - e^{l(\eta)}| < e^{l_n(\eta)} + e^{l(\eta)} < \varepsilon(1 + e^\varepsilon) \quad &\text{for } \eta \in \mathcal{E} \setminus F, \\ |e^{l_n(\eta)} - e^{l(\eta)}| < \varepsilon < \varepsilon(1 + e^\varepsilon) \quad &\text{for } \eta \in F. \end{aligned} \tag{2.47}$$

Since $\varepsilon(1 + e^\varepsilon) \to 0$ as $\varepsilon \to 0$, the proof follows.

□

### 2.2.2 Properties of the $\Delta_P$ function

The function $\Delta_P$, given by (2.15), consists of two summands, shown below

$$\mathcal{T}_P(\mathcal{A}) := \sum_{G\in\mathcal{A}} P(G) \cdot \mathsf{A}^*\big(E_P(T(X) \,|\, X \in G)\big) \quad \text{and} \quad \mathcal{H}_P(\mathcal{A}) := \sum_{G\in\mathcal{A}} P(G)\log P(G) \tag{2.48}$$

**Definition 2.20.** Let $\mathcal{A}$ and $\mathcal{B}$ be two partitions of the same space $\mathcal{X}$. We say that $\mathcal{A}$ is *finer* than $\mathcal{B}$ if for every $A \in \mathcal{A}$ there exist $B \in \mathcal{B}$ such that $A \subseteq B$. We can also say that $\mathcal{B}$ is *coarser* that $\mathcal{A}$. In such case we write $\mathcal{A} \preceq \mathcal{B}$; this relation is a partial order in the space of all partitions of $\mathcal{X}$. We use the same definition with respect to $P$-partitions of $\mathcal{X}$ (which are not necessarily partitions).

**Lemma 2.21.** *The function $\mathcal{T}_P(\mathcal{A})$ is decreasing with respect to partial order $\preceq$ on the space of finite $P$–partitions of $\mathbb{R}^d$.*

*Proof.* Let $\mathcal{A}, \mathcal{B}$ be two finite $P$-partitions of $\mathbb{R}^d$ and let $\mathcal{A} \preceq \mathcal{B}$. By an easy induction argument it is enough to show that for $A, B \in \mathcal{A}$

$$P(A)\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) + P(B)\mathsf{A}^*\big(E_P(T(X) \,|\, X \in B)\big) \geq P(C)\mathsf{A}^*\big(E_P(T(X) \,|\, X \in C)\big) \tag{2.49}$$

where $C = A \cup B$. The log-partition function $\mathsf{A}$ is convex (Theorem 1.2) and hence its convex conjugate $\mathsf{A}^*$ is also convex (Rockafellar, 1970, Theorem 12.2). Therefore

$$\begin{aligned} &\frac{P(A)}{P(C)}\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big) + \frac{P(B)}{P(C)}\mathsf{A}^*\big(E_P(T(X) \,|\, X \in B)\big) \geq \\ &\quad \geq \mathsf{A}^*\Big(\frac{P(A)}{P(C)}E_P(T(X) \,|\, X \in A) + \frac{P(B)}{P(C)}E_P(T(X) \,|\, X \in B)\Big) = \\ &\quad = \mathsf{A}^*\big(E_P(T(X) \,|\, X \in C)\big) \end{aligned} \tag{2.50}$$

and that concludes the proof. □

**Lemma 2.22.** *The function $\mathcal{H}_P(\mathcal{A})$ is increasing with respect to partial order $\preceq$ on the space of finite $P$–partitions of $\mathbb{R}^d$.*

*Proof.* Let $\mathcal{A}, \mathcal{B}$ be two finite $P$-partitions of $\mathbb{R}^d$ and let $\mathcal{A} \preceq \mathcal{B}$. By an easy induction argument it is enough to show that for $A, B \in \mathcal{A}$

$$P(A)\log P(A) + P(B)\log P(B) \leq P(C)\log P(C). \tag{2.51}$$

We have

$$P(A)\log P(A) + P(B)\log P(B) - P(C)\log P(C) = P(A)\log\frac{P(A)}{P(C)} + P(B)\log\frac{P(B)}{P(C)} \leq 0 \tag{2.52}$$

and the proof follows. The last inequality in (2.52) comes from $P(A), P(B) \leq P(C)$. □

Lemma 2.21 and Lemma 2.22 motivate the usage of the symbol '$\Delta_P$' – it represents a function being a difference of two other functions, both decreasing with respect to partition order. Hence optimizing the $\Delta_P$ function can be thought as finding a balance between coarse and fine partitions.

**Remark 2.23.** Let $X \sim P$ be a random vector with values in the observation space $\mathcal{X}$ and let $\mathcal{A}$ be a $P$-partition of the observation space. Let $X^{-1}(\mathcal{A}) = \{X^{-1}(A) \colon A \in \mathcal{A}\}$. Then we can write $\mathcal{T}_P(\mathcal{A})$ more succinctly (and perhaps more artificially) as

$$\mathcal{T}_P(\mathcal{A}) = E_P\mathsf{A}^*\big(E_P(T(X) \,|\, \sigma(X^{-1}(\mathcal{A})))\big). \tag{2.53}$$

Recall that convex conjugate $\mathsf{A}^*$ of a convex function $A$ is a convex function itself (Rockafellar, 1970). Therefore by (2.53) and Jensen's inequality

$$\mathcal{T}_P(\mathcal{A}) \leq E_P\Big(E_P(\mathsf{A}^*\big(T(X)\big) \,|\, \sigma(X^{-1}(\mathcal{A})))\Big) = E_P(\mathsf{A}^*\big(T(X)\big)) \tag{2.54}$$

Note that it is possible that $E_P\mathsf{A}^*\big(T(X)\big) = \infty$ (in particular, it may be the case that $\mathsf{A}^*\big(T(\boldsymbol{x})\big) = \infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$, as is the case with the Normal-Inverse-Wishart or Normal-Inverse-Gamma models). However, if $E_P\mathsf{A}^*\big(T(X)\big) < \infty$, then (2.54) gives an upper bound on $\mathcal{T}_P(\mathcal{A})$, and in turn an upper bound on $\Delta_P(\mathcal{A})$ (since $\mathcal{H}_P$ is a non-positive function).

### 2.2.3 The $\Delta_P$ function in the Gaussian case

In this section we compute the exact formula for the $\Delta_P$ function in Bayesian Mixture Models with conjugate Normal priors, presented in Section 1.4.1. It should be noted that in these cases this formula could be computed by a direct calculation on the exact formula for the likelihood. For example, for the Normal-Normal model this direct approach was applied in Rajkowski, 2019, Lemma 4.5 (for a special case of the Chinese Restaurant Process prior on the space of partitions). However here we want to apply the more general formula from Theorem 2.12 and hence our goal is simply to compute the value of $\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big)$ for these model specifications. We also check that the assumptions (i) and (ii) of Theorem 2.12 hold for these models if $P$ is a distribution continuous with respect to the Lebesgue measure (the regularity and the convexity assumptions were established in Section 1.4.1).

The following standard result from the theory of exponential families will be useful. It is a straightforward consequence of Theorem 1.6.4, Theorem 2.3.1 and Corollary 2.3.1 in Bickel and Doksum (2015).

**Theorem 2.24.** *Suppose $\mathcal{P}$ is the canonical exponential family generated by $(T, h)$ and that*

*(i) The natural parameter space, $\mathcal{E}$, is open,*

*(ii) $\eta$ is identifiable.*

*Let $x$ be the observed data vector and set $t_0 = T(x)$. If $C_T$ is the convex support of the distribution $T(X)$ then $\hat{\eta} = \operatorname{argmax}_{\eta\in\mathcal{E}} \left(\eta^\top T(x) - \mathsf{A}(\eta)\right)$ exists and is unique if and only if $t_0 \in C_T^0$ where $C_T^0$ is the interior of $C_T$. In such case $\hat{\eta}$ satisfies*

$$\mathbb{E}_{\hat{\eta}}\big(T(X)\big) = t_0. \tag{2.55}$$

**Corollary 2.25.** *With the assumptions of Theorem 2.24 we have $\mathsf{A}^*(t_0) = {t_0}^\top\hat{\eta} - \mathsf{A}(\hat{\eta})$, where $\hat{\eta}$ satisfies $\mathbb{E}_{\hat{\eta}}T(\boldsymbol{x}) = t_0$.*

**Note 2.26.** If $P$ is a probability distribution continuous with respect to the Lebesge measure, $A \subset \mathbb{R}^d$ is a $P$-measurable set and $T(A)$ is not a singleton then $E_P(T(X) \,|\, X \in$

$A) \in C_T^0$ and hence the assumption $(i)$ of Theorem 2.12 is satisfied. This property is easily verified by all the Normal models given below, hence we check only the assumption (ii) of Theorem 2.12.

<u>**Normal-Normal**</u>

**Assumption (ii) of Theorem 2.12.** By investigation of (1.32) we have that $[\tau_0, \zeta_0](\mu_0, \Psi_0) = [\tau_0(\mu_0, \Psi_0)^\top, \zeta_0(\mu_0, \Psi_0)^\top]^\top \in \Omega$ for every $(\mu_0, \Psi_0) \in \mathbb{R}^d \times \mathcal{S}_+^d$. It is clear that $[\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d)$ is an open subset of $\Omega$ and that for every $P$-measurable $A$ we have $\big(E(T(X) \,|\, X \in A), \boldsymbol{a}\big) \in [\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d)$, and hence our assumption is satisfied for $r = 1$.

**Computation of $\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big)$.** The computation of $\mathsf{A}^*\big(E(T(X) \,|\, X \in A)\big)$ is also straightforward since $\eta = \theta = \mu$ is the natural parameter. Let us recall (1.31) from which $T(x) = \Sigma_0^{-1} x$ and hence

$$\mathbb{E}_\mu T(x) = \Sigma_0^{-1}\mu. \tag{2.56}$$

Let $t_0 = E_P(T(X) \,|\, X \in A)$. By Theorem 2.24, $\Sigma_0^{-1}\hat{\mu} = E_P(\Sigma_0^{-1} X \,|\, X \in A)$. By the linearity of the expected value this implies

$$\hat{\mu} = E_P(X \,|\, X \in A). \tag{2.57}$$

Using (2.57) it is straightforward to verify that

$$t_0^\top \hat{\mu} = \hat{\mu}\Sigma_0^{-1}\hat{\mu}. \tag{2.58}$$

By (1.31) we have $\mathsf{A}(\eta) = \mathsf{B}(\theta) = \frac{1}{2}\mu^\top \Sigma_0^{-1}\mu$ and hence, using (2.58) and Corollary 2.25

$$\mathsf{A}^*(t_0) = \hat{\mu}\Sigma_0^{-1}\hat{\mu} - \frac{1}{2}\hat{\mu}\Sigma_0^{-1}\hat{\mu} = \frac{1}{2}\hat{\mu}\Sigma_0^{-1}\hat{\mu} = \frac{1}{2}\|E_P(RX \,|\, X \in A)\|^2, \tag{2.59}$$

where $R^2 = \Sigma_0^{-1}$. Therefore in the Normal-Normal case the $\Delta_P$ function is equal to

$$\Delta_P^{NN}(\mathcal{A}) = \frac{1}{2}\sum_{A \in \mathcal{A}} P(A) \cdot \|E_P(RX \,|\, X \in A)\|^2 + \mathcal{H}(\mathcal{A}). \tag{2.60}$$

Note that

$$\|E_P(RX \,|\, X \in A)\|^2 = E_P(\|RX\|^2 \,|\, X \in A) - \mathrm{tr}\mathbf{V}_P(RX \,|\, X \in A) \tag{2.61}$$

and

$$\sum_{A \in \mathcal{A}} P(A) E_P(\|RX\|^2 \,|\, X \in A) = E_P(\|RX\|^2), \tag{2.62}$$

hence in this case we can reformulate $\Delta_P^{NN}$ to

$$\Delta_P^{NN}(\mathcal{A}) = \frac{1}{2}\mathbb{E}_P(\|RX\|^2) - \frac{1}{2}\sum_{A \in \mathcal{A}} P(A) \cdot \mathrm{tr}\mathbf{V}_P(RX \,|\, X \in A) + \mathcal{H}(\mathcal{A}). \tag{2.63}$$

It then follows that

$$\mathcal{T}_P(\mathcal{A}) = \frac{1}{2}\mathbb{E}_P(\|RX\|^2) - \frac{1}{2}\sum_{A \in \mathcal{A}} P(A) \cdot \mathrm{tr}\mathbf{V}_P(RX \,|\, X \in A) \leq \frac{1}{2}\mathbb{E}_P(\|RX\|^2), \tag{2.64}$$

which could be also obtained from the more general form (2.54).

**Normal-Inverse-Wishart**

**Assumption (ii) of Theorem 2.12.** The hyperparameter space for $(\mu_0, \Sigma_0, \nu_0, \kappa_0)$ is $\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+$. The image $[\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+)$ is open in $\Omega$. If $P$ is a continuous distribution and $P(A) > 0$ then $E_P(XX^\top \,|\, X \in A)$ is positive definite matrix. Then for $r \geq 2d+4$ we can express $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big)$ as $[\tau_0, \zeta_0](\mu_0, \Sigma_0, \nu_0, \kappa_0)$ for appropriately chosen $(\mu_0, \Sigma_0, \nu_0, \kappa_0) \in \mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+$ (the assumption that $r \geq 2d+4$ is important due to the first coordinate of $\zeta_0$). In other words, $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big) \in [\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+)$, and the proof follows.

**Computation of $\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big)$.** Recall that the parameter space is $\Theta = \mathbb{R}^d \times \mathcal{S}_+^d$ and it bijectively corresponds to the natural parameter space $\mathcal{E} \subseteq \mathbb{R}^{\frac{d(d+3)}{2}}$, where the bijection is given by $\eta(\theta) = \eta\big(\mu, \Lambda\big) = \big(\mathrm{diag}(\Lambda^{-1}), \mathrm{low}(\Lambda^{-1}), \Lambda^{-1}\mu\big)$. (See the beginning of Section 1.4.1 for the definition of 'diag' and 'low'.) Let $x \sim \mathcal{N}(\mu, \Lambda^{-1})$. By the well known properties of the Normal distribution

$$\begin{aligned} \mathbb{E}_\eta x &= \mu \\ \mathbb{E}_\eta xx^\top &= \Lambda^{-1} + \mu\mu^\top \end{aligned} \tag{2.65}$$

Using the formula for $T(x)$ in the NIW model (cf. (1.41)) and the linearity of the expected value and low and diag operations:

$$\mathbb{E}_\eta T(x) = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\mathbb{E}_\eta xx^\top) \\ -\mathrm{low}(\mathbb{E}_\eta xx^\top) \\ \mathbb{E}_\eta x \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\Lambda + \mu\mu^\top) \\ -\mathrm{low}(\Lambda + \mu\mu^\top) \\ \mu \end{bmatrix}. \tag{2.66}$$

Let $A \subseteq \mathbb{R}^d$ be a $P$-measurable set and let $t_0 = E_P(T(X) \,|\, X \in A)$. Let $\hat{\theta} = (\hat{\mu}, \hat{\Lambda}) = \theta(\hat{\eta})$. By Corollary 2.25 and the formula for $\mathsf{A}(\eta) = \mathsf{B}(\theta(\eta))$ in (1.41)

$$\mathsf{A}^*(t_0) = t_0\hat{\eta} - \mathsf{A}(\hat{\eta}) = t_0\hat{\eta}^\top - \frac{1}{2}\log|\hat{\Lambda}| - \frac{1}{2}\hat{\mu}\hat{\Lambda}^{-1}\hat{\mu} \tag{2.67}$$

Note that Corollary 2.25 we have $t_0 = \mathbb{E}_{\hat{\eta}} T(x)$ and hence by (2.66)

$$\begin{aligned} t_0{}^\top\hat{\eta} &= \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top) \\ -\mathrm{low}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top) \\ \hat{\mu} \end{bmatrix}^\top \begin{bmatrix} \mathrm{diag}(\hat{\Lambda}^{-1}) \\ \mathrm{low}(\hat{\Lambda}^{-1}) \\ \hat{\Lambda}^{-1}\hat{\mu} \end{bmatrix} = \\ &= -\frac{1}{2}\mathrm{diag}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top)^\top\mathrm{diag}(\hat{\Lambda}^{-1}) - \mathrm{low}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top)^\top\mathrm{low}(\hat{\Lambda}^{-1}) + \hat{\mu}^\top\hat{\Lambda}^{-1}\hat{\mu} = \\ &= -\frac{1}{2}\mathrm{tr}(\hat{\Lambda}\hat{\Lambda}^{-1}) - \frac{1}{2}\mathrm{tr}(\hat{\mu}\hat{\mu}^\top\hat{\Lambda}^{-1}) + \hat{\mu}\hat{\Lambda}^{-1}\hat{\mu} = -\frac{d}{2} + \frac{1}{2}\hat{\mu}\hat{\Lambda}^{-1}\hat{\mu}. \end{aligned} \tag{2.68}$$

Joining (2.67) with (2.68) we obtain

$$\mathsf{A}^*(t_0) = -\frac{d}{2} - \frac{1}{2}\log|\hat{\Lambda}| \tag{2.69}$$

We now write the relationship $E_P(T(X) \,|\, X \in A) = t_0 = \mathbb{E}_{\hat{\eta}} T(X)$ using (2.66) and the formula for $T(x)$ (cf. (1.41)):

$$\begin{bmatrix} -\frac{1}{2}\mathrm{diag}(E_P(XX^\top \,|\, X \in A)) \\ -\mathrm{low}(E_P(XX^\top \,|\, X \in A)) \\ E_P(X \,|\, X \in A) \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top) \\ -\mathrm{low}(\hat{\Lambda} + \hat{\mu}\hat{\mu}^\top) \\ \hat{\mu} \end{bmatrix}. \tag{2.70}$$

From (2.70) we deduce that

$$\hat{\Lambda} = E_P(XX^\top \,|\, X \in A) - E_P(X \,|\, X \in A)E_P(X \,|\, X \in A)^\top = \mathbf{V}_P(X \,|\, X \in A) \tag{2.71}$$

and hence, by (2.69)

$$\mathsf{A}^*(t_0) = -\frac{d}{2} - \frac{1}{2}\log|\mathbf{V}_P(X \,|\, X \in A)| \tag{2.72}$$

Plugging this to (2.48) we get that for Normal-Inverse-Wishart model

$$\begin{aligned}\Delta_P^{NIW}(\mathcal{A}) &= -\frac{1}{2}\sum_{A\in\mathcal{A}} P(A)\cdot\big(d + \log|\mathbf{V}_P(X \,|\, X \in A)|\big) + \mathcal{H}(\mathcal{A}) = \\ &= -\frac{d}{2} - \frac{1}{2}\sum_{A\in\mathcal{A}} P(A)\cdot\log|\mathbf{V}_P(X \,|\, X \in A)| + \mathcal{H}(\mathcal{A})\end{aligned} \tag{2.73}$$

Note that for any $x \in \mathbb{R}^d$ we have

$$\mathsf{A}^*(T(x)) = \sup_{(\mu,\Lambda)\in\mathbb{R}^d\times\mathcal{S}_+^d}\Big(-\frac{1}{2}(x-\mu)^\top\Lambda^{-1}(x-\mu) - \frac{1}{2}\log|\Lambda|\Big) = \infty, \tag{2.74}$$

since we can take $\mu = x$ and $\Lambda$ of arbitrarily small norm. Hence the bound (2.54) for $\mathcal{T}_P$ cannot be applied in this case.

<u>**Normal-Inverse-Gamma**</u>

**Assumption (ii) of Theorem 2.12.** The hyperparameter space for $(\mu_0, \Psi_0, \beta_0, \gamma_0)$ is $\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+$. The image $[\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+)$ is open in $\Omega$. For $r \geq 2$ we can express $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big)$ as $[\tau_0, \zeta_0](\mu_0, \Psi_0, \beta_0, \kappa_0)$ for appropriately chosen $(\mu_0, \Psi_0, \beta_0, \gamma_0) \in \mathbb{R}^d\times\mathcal{S}_+^d\times\mathbb{R}_+\times\mathbb{R}_+$ (the assumption that $r \geq 2$ is important due to the first coordinate of $\zeta_0$). In other words, $\big(rE_P(T(X) \,|\, X \in A), r\boldsymbol{a}\big) \in [\tau_0, \zeta_0](\mathbb{R}^d \times \mathcal{S}_+^d \times \mathbb{R}_+ \times \mathbb{R}_+)$, and the proof follows.

**Computation of $\mathsf{A}^*\big(E_P(T(X) \,|\, X \in A)\big)$.** Similarly to the case of the Normal-Normal model, let us use $R$ to denote the symmetric matrix that satisfies $R^2 = \Sigma_0^{-1}$. Note that in the Normal-Inverse-Gamma model, in which the distribution of $x$ (given $\theta = (\mu, \lambda)$) is $\mathcal{N}(\mu, \lambda\Sigma_0)$, we have

$$\begin{aligned}\mathbb{E}_\eta x^\top\Sigma_0^{-1}x = \mathbb{E}_\eta\|Rx\|^2 &= \mathrm{tr}\big(\mathbf{V}_\eta(Rx)\big) + \|\mathbb{E}_\eta Rx\|^2 = \\ &= \mathrm{tr}(\lambda I_d) + \|R\mu\|^2 = d\lambda + \|R\mu\|^2.\end{aligned} \tag{2.75}$$

This, together with the formula for the sufficient statistic $T(x)$ in this model (cf. (1.53)) implies

$$\mathbb{E}_\eta T(x) = \begin{bmatrix} -\frac{1}{2}\mathbb{E}_\eta x^\top\Sigma_0^{-1}x \\ \Sigma_0^{-1}\mathbb{E}_\eta x\end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(d\lambda + \|R\mu\|^2) \\ \Sigma_0^{-1}\mu\end{bmatrix}. \tag{2.76}$$

Let $t_0 = E_P(T(X) \,|\, X \in A)$. By Theorem 2.24 we have $t_0 = \mathbb{E}_{\hat{\eta}}T(x)$, which leads to

$$\begin{bmatrix} -\frac{1}{2}E_P(X^\top\Sigma_0^{-1}X \,|\, X \in A) \\ \Sigma_0^{-1}\mathbb{E}_P(X \,|\, X \in A)\end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(d\hat{\lambda} + \|R\hat{\mu}\|^2) \\ \Sigma_0^{-1}\hat{\mu}\end{bmatrix}. \tag{2.77}$$

in which $(\hat{\mu}, \hat{\lambda}) = \theta(\hat{\eta})$. By (2.77) we get $\hat{\mu} = E_P(X \mid X \in A)$ and

$$-\frac{d}{2}\hat{\lambda} = -\frac{1}{2}E_P(\|RX\|^2 \mid X \in A) + \frac{1}{2}\|R\hat{\mu}\|^2 = -\frac{1}{2}\mathrm{tr}\mathbf{V}_P(RX \mid X \in A). \tag{2.78}$$

Note that

$$\begin{aligned} {t_0}^\top \hat{\eta} &= \begin{bmatrix} -\frac{1}{2}(d\hat{\lambda} + \|R\hat{\mu}\|^2) \\ \Sigma_0^{-1}\hat{\mu} \end{bmatrix}^\top \begin{bmatrix} 1/\hat{\lambda} \\ \hat{\mu}/\hat{\lambda} \end{bmatrix} = \\ &= -\frac{d}{2} - \frac{1}{2}\|R\hat{\mu}\|^2/\hat{\lambda} + \|R\hat{\mu}\|^2/\hat{\lambda} = -\frac{d}{2} + \frac{1}{2}\|R\hat{\mu}\|^2/\hat{\lambda} \end{aligned} \tag{2.79}$$

and hence, by Corollary 2.25, the formula for $\mathsf{B}(\theta)$ in (1.53) and (2.78)

$$\begin{aligned} \mathsf{A}^*(t_0) &= \Big(-\frac{d}{2} + \frac{1}{2}\|R\hat{\mu}\|^2/\hat{\lambda}\Big) - \Big(\frac{d}{2}\log\hat{\lambda} + \frac{1}{2}\hat{\mu}^\top\Sigma_0^{-1}\hat{\mu}/\hat{\lambda}\Big) = -\frac{d}{2} - \frac{d}{2}\log\hat{\lambda} = \\ &= -\frac{d}{2} - \frac{d}{2}\big(\log\mathrm{tr}\mathbf{V}_P(RX \mid X \in A)\big) - \log d \end{aligned} \tag{2.80}$$

which leads us to

$$\Delta_P^{NIG}(\mathcal{A}) = -\frac{d}{2}\big(1 - \log d\big) - \frac{1}{2}\sum_{A\in\mathcal{A}} P(A)\cdot\log\mathrm{tr}\mathbf{V}_P(RX \mid X \in A) + \mathcal{H}(\mathcal{A}). \tag{2.81}$$

Note that, similarly as in the Inverse-Wishart case, for any $x \in \mathbb{R}^d$ we have

$$\mathsf{A}^*(T(x)) = \sup_{(\mu,\lambda)\in\mathbb{R}^d\times\mathbb{R}_+}\Big(-\frac{1}{2\lambda}(x-\mu)^\top\Sigma_0^{-1}(x-\mu) - \frac{d}{2}\log\lambda\Big) = \infty, \tag{2.82}$$

since we can take $\mu = x$ and arbitrarily small $\lambda$. Hence the bound (2.54) for $\mathcal{T}_P$ cannot be applied in this case.

### Maximisation of $\Delta_P$ in the Normal when $P = \mathrm{Unif}([0,1])$

The $\Delta_P$ function was derived as a asymptotic limit of the logarithm of the BMM posterior score of an induced partition, when the data is independently sampled from some distribution $P$. In our quest of describing the MAP clustering, it therefore seems informative to consider partitions of the observation space that yield the highest posterior score – which necessarily are also the maximisers of the $\Delta_P$ function, as is easily deduced from Corollary 2.13. For the Gaussian models, a computationally tractable example, in which we are able to describe the maximiser exactly (or prove that it does not exist), is the case when the input distribution $P$ is the uniform distribution on the segment $[0,1]$. Note that in the one-dimensional case the Normal-Inverse-Wishart and Normal-Inverse-Gamma models are equivalent and hence we will consider only the former there.

We start with Lemma 2.27 which in the case of the uniform input distribution $P$ will allow us to narrow the search for the maximiser of the $\Delta_P$ function in the conjugate Normal models to the divisions of $[0,1]$ into subsegments. Basically it says that when a distribution $P$ is continuous with respect to the Lebsgue measure on a line, then segments are those sets that minimise the variance given their probability (which is quite natural and expected).

**Lemma 2.27.** *Let $P$ have a density with respect to the Lebesgue measure on $\mathbb{R}$. Let $A$ by any $P$-measurable set and let $S_A$ be a closed segment (i.e. a set of the form $[a, b]$ for $a < b$) centered at $e_A := E_P(X \mid X \in A)$ such that $P(S_A) = P(A)$. Then $\mathrm{Var}_P(X \mid X \in A) \geq \mathrm{Var}_P(X \mid X \in S_A)$ and the equality holds only if $P(A \setminus S_A) = P(S_A \setminus A) = 0$.*

*Proof.* Note that the existence of $S_A$ is guaranteed by the continuity of $P$. Let $r_A$ be the radius of $S_A$ (i.e. $r_A$ is half the length of segment $S_A$). We have

$$\mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in A \setminus S_A} \geq \mathbb{E}\, r_A^2 \mathbf{1}_{X \in A \setminus S_A} = r_A^2 P(A \setminus S_A). \tag{2.83}$$

with the equality only if $P(A \setminus S_A) = 0$. Similarly

$$\mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in S_A \setminus A} \leq \mathbb{E}\, r_A^2 \mathbf{1}_{X \in S_A \setminus A} = r_A^2 P(S_A \setminus A). \tag{2.84}$$

Since $P(A) = P(S_A)$ we have that $P(A \setminus S_A) = P(S_A \setminus A)$ and hence, by (2.84) and (2.83) we get

$$\mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in A \setminus S_A} \geq \mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in S_A \setminus A}. \tag{2.85}$$

Adding $\mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in A \cap S_A}$ to the both sides of (2.85) yields

$$\mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in A} \geq \mathbb{E}\,(X - e_A)^2 \mathbf{1}_{X \in S_A} \geq \mathbb{E}\,(X - e_{S_A})^2 \mathbf{1}_{X \in S_A}, \tag{2.86}$$

where $e_{S_A} := \mathbb{E}\,(X \mid X \in S_A)$. By dividing (2.86) by $P(A) = P(S_A)$ we get $\mathrm{Var}_P(X \mid X \in A) \geq \mathrm{Var}_P(X \mid X \in S_A)$ (with the equality only if $P(A \setminus S_A) = 0$). This finishes the proof of Lemma 2.27. □

The condition $P(A \setminus S_A) = P(S_A \setminus A) = 0$ is equivalent to $d_P(A, S_A) = 0$, where $d_P$ is the *symmetric distance pseudometric*, defined below.

**Definition 2.28.** Let $\mathcal{M}$ be a $\sigma$-field on $\mathbb{R}^d$ and $\mu$ be a measure on $(\mathbb{R}^d, \mathcal{M})$. Then the function $d_\mu \colon \mathcal{M}^2 \to \mathbb{R}$ defined by $d_\mu(A, B) = \mu\big((A \setminus B) \cup (B \setminus A)\big)$ 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 $d_\mu$ 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 following Corollary 2.29 states that the only possible maximisers of the $\Delta_P$ function in the Gaussian case (when $P$ is the uniform distribution on $[0, 1]$) are basically (up to $d_P$ distance) the divisions of $[0, 1]$ into subsegments.

**Corollary 2.29.** *Consider the function $\Delta_P(\mathcal{A})$ given by (2.73) or (2.63) and let $P$ be the uniform distribution on the segment $[0, 1]$. If $\hat{\mathcal{A}}$ is any maximiser of $\Delta_P$ among finite $P$-partitions of the segment $[0, 1]$ then $d_P(A, S_A) = 0$ for any $A \in \hat{\mathcal{A}}$, where $S_A$ is defined as in the formulation of Lemma 2.27.*

*Proof.* Assume that $\hat{\mathcal{A}} = \{A_1, \ldots, A_K\}$ is the maximiser of the function $\Delta_P$ among all finite $P$-partitions of $[0,1]$. Let us extend the domain of the function $\Delta_P$ not only to finite $P$-partitions of the segment $[0,1]$, but also to simply finite collections of $P$-measurable subsets of $[0,1]$. Let $\mathcal{S} = \{S_{A_i} : i = 1, \ldots, K\}$. Since for every $A \in \hat{\mathcal{A}}$ we have $P(S_A) = P(A)$ and $\mathrm{Var}_P(S_A) \le \mathrm{Var}_P(A)$, it follows from the formulas (2.73) and (2.63) that in these cases $\Delta_P(\hat{\mathcal{A}}) \le \Delta_P(\mathcal{S})$. Now let $\mathcal{S}' = \{S_1, \ldots, S_K\}$ be any partition of the segment $[0,1]$ into subsegments such that $P(S_i) = P(S_{A_i})$ (i.e. the lengths of segments $S_i$ and $S_{A_i}$ are the same). Such partition exists, since $\sum_{i=1}^K P(S_{A_i}) = \sum_{i=1}^K P(A_i) = 1$. By the properties of the uniform distribution we also have $\mathrm{Var}_P(X \,|\, X \in S_i) = \mathrm{Var}_P(X \,|\, X \in S_{A_i})$. It follows that in the Gaussian case $\Delta_P(\mathcal{S}') = \Delta_P(\mathcal{S}) \ge \Delta_P(\hat{\mathcal{A}})$. On the other hand, $\hat{\mathcal{A}}$ is the maximiser of $\Delta_P$, so $\Delta_P(\mathcal{S}') = \Delta_P(\hat{\mathcal{A}})$. Using the condition for the equality in Lemma 2.27, we get $d_P(A, S_A) = 0$ for every $A \in \hat{\mathcal{A}}$. □

**Normal-Normal model.** Let $\mathcal{S} = \{S_1, \ldots, S_K\}$ be a partition of $[0,1]$ into subsegments of length $p_1, \ldots, p_K$ respectively. Note that $\mathrm{Var}_P(X \,|\, X \in S_i) = p_i^2 v$, where $v = \frac{1}{12}$ is the total variance in the segment, and $E_P(\|RX\|^2) = R^2 E_P X^2 = \frac{1}{3}$. It follows from (2.63) that

$$\Delta_P^{NN}(\mathcal{S}) = \frac{R^2}{6} - \frac{1}{2}\sum_{i=1}^K p_i \cdot R^2 \cdot p_i^2 v + \sum_{i=1}^K p_i \log p_i. \tag{2.87}$$

**Proposition 2.30.** *The values of $K$ and $p_1, \ldots, p_K$ (such that $\sum_{i=1}^K p_i = 1$) that maximise (2.87) satisfy*

$$\hat{K} \in \{\lfloor R^2 v \rfloor, \lceil R^2 v \rceil\}, \quad \hat{p}_1 = \ldots = \hat{p}_{\hat{K}} = \frac{1}{\hat{K}} \tag{2.88}$$

*(where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and the ceiling functions respectively).*

*Proof.* Although at first sight this seems to be a standard problem solved by the Lagrange multipliers technique, such approach has some subtle difficulties. To circumvent them carefully, we will optimize (2.87) 'locally', focusing on two segments at a time, leaving others unchanged.

Let $F_K(p_1, \ldots, p_K)$ be the right-hand side of (2.87). We now prove that we increase $F_K$ by replacing arguments $(p_1, p_2)$ by at least one of $(0, p_1 + p_2)$ or $\left(\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}\right)$. More precisely, we show that

if $p_1 \ne p_2$ and $p_1 p_2 \ne 0$ then

$$F_K(p_1, \ldots, p_K) < \max\left\{F_K(0, p_1 + p_2, p_3, \ldots, p_K), F\Big(\frac{p_1 + p_2}{2}, \frac{p_1 + p_2}{2}, p_3, \ldots, p_K\Big)\right\}. \tag{2.89}$$

Let us consider $p_1, \ldots, p_K$ as fixed and let $f(p) = F_K(p, p_1 + p_2 - p, p_3, \ldots, p_K)$, where $0 \le p \le p_1 + p_2$. For simplicity of notation, let $q = p_1 + p_2 - p$. By direct calculation

$$f'(p) = -\frac{3R^2 v}{2}\big(p^2 - q^2\big) + \log p - \log q. \tag{2.90}$$

It follows that

$$\lim_{p\to 0} f'(p) = -\infty, \quad \lim_{p\to p_1+p_2} f'(p) = \infty, \quad f'\left(\frac{p_1+p_2}{2}\right) = 0. \tag{2.91}$$

Again, direct calculation leads

$$f''(p) = -3R^2 v + \frac{1}{p} + \frac{1}{q}, \quad f'''(p) = -\frac{1}{p^2} + \frac{1}{q^2}. \tag{2.92}$$

Hence $f'''(p) = 0$ implies $p = q = \frac{p_1+p_2}{2}$, i.e. there is only one root of $f'''(p) = 0$ on $(0, p_1+p_2)$. By applying Rolle's theorem twice we get that there are at most two roots of $f''(p) = 0$ in $(0, p_1+p_2)$ and, in turn,

$$\text{there are at most three roots of } f'(p) = 0 \text{ in } (0, p_1+p_2). \tag{2.93}$$

Now, from (2.91), (2.93) and standard one-variate analysis we deduce that $f(p)$ achieves its maximal value for $p = 0$, $p = p_1 + p_2$ or $p = \frac{p_1+p_2}{2}$. By translating this into function $F_K$, we get (2.89).

Let us come back to the initial problem. The function $F_K$ is a continuous function on the compact set $\triangle^K = \{(p_1, \ldots, p_K) \colon \sum_{i=1}^K p_i = 1, \forall_{i\le K} p_i \ge 0\}$. Therefore it achieves its maximal value for some $(\hat{p}_1 \ldots, \hat{p}_K) \in \triangle^K$. It follows from (2.89) and the symmetry of $F_K$ that there exist $I \subseteq [n]$ such that $\hat{p}_i = \frac{1}{|I|}$ for $i \in I$ and $\hat{p}_i = 0$ for $i \notin I$. Note that

$$F_K(\hat{p_1}, \ldots, \hat{p_K}) = F_{|I|}(\frac{1}{|I|}, \ldots, \frac{1}{|I|}) \tag{2.94}$$

and therefore it is left to maximise $F_K\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$ for $K \in \mathbb{N}$. We have

$$F_K\left(\frac{1}{K}, \ldots, \frac{1}{K}\right) = \frac{R^2}{6} - \frac{R^2 v}{2}\left(\frac{1}{K}\right)^2 + \log\frac{1}{K}. \tag{2.95}$$

The function $x \mapsto \frac{R^2}{6} - \frac{R^2 v}{2x^2} - \log x$ has its only extreme value in $x_0 = \sqrt{R^2 v}$, which is easily established by calculation of the derivative. Hence the optimal number of clusters is $\hat{K} = \lfloor x_0 \rfloor$ or $\hat{K} = \lceil x_0 \rceil$. □

It is worth pointing out that if $x_0 \in \mathbb{N}$ then the variance of every segment becomes simply $\frac{1}{R^2} = \Sigma_0$. It can be said that in this case the optimal partition 'adapts' itself to the covariance hyperparameter $\Sigma_0$.

**Normal Inverse-Wishart** Let $\mathcal{S} = \{S_1, \ldots, S_K\}$ be a partition of $[0, 1]$ into subsegments of length $p_1, \ldots, p_K$ respectively. Recall that $\mathrm{Var}_P(X \mid X \in S_i) = p_i^2 v$, where $v = \frac{1}{12}$ is the total variance in the segment. It follows from (2.73) that

$$\Delta_P^{NIW}(\mathcal{S}) = -\frac{d}{2} - \frac{1}{2}\sum_{i\le n} p_i \log(p_i^2 v) + \sum_{i\le n} p_i \log p_i = \frac{\log(12)}{2}. \tag{2.96}$$

Hence every partition of $[0, 1]$ into subsegments gives the same $\Delta_P^{NIW}$ score. This shows that the maximiser is not unique in this case and, indeed, we have constructed an infinite family of maximisers; we can find a maximiser with arbitrarily many clusters.

The result is quite intuitive; with no prior suggestion of preference, the classifier should have no reason to express preference between a single cluster $[0, 1]$ and two clusters, $[0, 0.5]$ and $[0.5, 1]$.

# Chapter 3

# Asymptotic Results for MAP clustering in the Normal-Normal BMM

In Section 2.2 we showed the approximation to the logarithm of the joint probability $Q$ of clustering and data (formula (1.8)) in the conjugate exponential BMM, applied to a dataset sampled independently from some probability distribution $P$ on $\mathbb{R}^d$ and a clustering induced by some finite $P$-partition of the observation space. This led us to a formulation of a function $\Delta_P$ on the space of finite $P$-partitions of the observation space. Maximisers of this function are partitions that can induce 'best' clusterings in terms of the posterior score $Q$. This does not imply that finite $P$-partitions that maximise the $\Delta_P$ function will induce the MAP clustering. In the first place, we do not even know if the MAP clustering is induced by any partition of the observation space. Potentially this clustering can change significantly every time a new observation is registered. On the other hand, as showed in Section 2.1, the MAP clustering possesses nice geometric properties, namely the clusters are separated by the contour lines of the linear functional of the sufficient statistic in the model. This can give us a hope that the limit of the MAP clusterings exists and exhibits some of the properties of the induced partitions. This line of research was pursued in Rajkowski (2019) with some success, for the Normal-Normal BMM when the prior on the space of clusterings is the Chinese Restaurant Process, given by (1.14). In this case, when the input distribution $P$ has a bounded support, then we can connect the limits of the MAP clusterings to the maximisers of the $\Delta_P^{NN}$ function. This chapter presents the details of this result from Rajkowski (2019).

It is important to underline that for this chapter **we restrict our attention to the Normal-Normal BMM in which the prior on the space of clusterings is the Chinese Restaurant Process.** Here we note that for the Normal-Normal model the marginal density $g_{NN,k}$ given by (1.34) depends on the mean location hyperparameter $\mu_0$ only via the shift $x \mapsto x - \mu_0$. (This applies also to the two remaining Normal models from

Section 1.4.1, but this is not the concern of this chapter). Therefore if we are interested in the clustering properties of the MAP clusterings in this model, we can without loss of generality consider the case $\mu_0 = 0$. The properties for the general case follow by an easy transition. In this case, using the definitions (1.8) with (1.34) and (1.14) we can reformulate the score function $Q$ so that it will be more convenient for further analysis.

**Remark 3.1.** The conditional probability of partition $\mathcal{I}$ in the zero-mean Normal-Normal BMM model with the Chinese Restaurant Process prior on the space of partitions, given the observation vector $\boldsymbol{x} = (x_j)_{j=1}^n$, is proportional to

$$\tilde{Q}_{\boldsymbol{x}}(\mathcal{I}) := C^{|\mathcal{I}|} \prod_{I \in \mathcal{I}} \frac{|I|!}{|I|^{(d+2)/2} \det R_{|I|}} \cdot \exp\left\{ \frac{1}{2} \sum_{I \in \mathcal{I}} |I| \cdot \left\| R_{|I|}^{-1} R^2 \overline{\boldsymbol{x}_I} \right\|^2 \right\} \tag{3.1}$$

where $C = \alpha / \sqrt{\det \Psi_0}$, $R = \Sigma_0^{-1/2}$, $R_k = (\Sigma_0^{-1} + \Psi_0^{-1}/k)^{1/2}$ for $k \in \mathbb{N}$, $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^d$.

*Proof.* Firstly note that

$$\Psi_k^{-1} = k R_k^2 \quad \text{and hence} \quad \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} = R^2 \frac{R_k^{-2}}{k} R^2 \text{ and } \det \Psi_k = k^{-d} (\det R_k)^{-2}. \tag{3.2}$$

It follows from (1.34) that for $\mu_0 = 0$:

$$\begin{aligned} g_{NN,k}(\boldsymbol{x} \,|\, \mathcal{I}) &= (2\pi |\Sigma_0|)^{-dn/2} \frac{|\Psi_k|^{|\mathcal{I}|/2}}{|\Psi_0|^{|\mathcal{I}|/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n x_i^\top \Sigma_0^{-1} x_i \right\} \exp\left\{ \frac{1}{2} \sum_{I \in \mathcal{I}} |I|^2 \overline{\boldsymbol{x}_I}^\top \Sigma_0^{-1} \Psi_{|I|} \Sigma_0^{-1} \overline{\boldsymbol{x}_I} \right\} = \\ &= (2\pi |\Sigma_0|)^{-dn/2} \frac{|\Psi_k|^{|\mathcal{I}|/2}}{|\Psi_0|^{|\mathcal{I}|/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n \|R x_i\|^2 \right\} \exp\left\{ \frac{1}{2} \sum_{I \in \mathcal{I}} |I| \|R_{|I|}^{-1} R^2 \overline{\boldsymbol{x}_I}\|^2 \right\} = \end{aligned} \tag{3.3}$$

From (1.14) we can write

$$\mathcal{P}_{\pi,n}(\mathcal{I}) \propto \alpha^{|\mathcal{I}|} \prod_{I \in \mathcal{I}} (|I| - 1)! \tag{3.4}$$

Now $\tilde{Q}_{\boldsymbol{x}}(\mathcal{I}) \propto Q(\boldsymbol{x}, \mathcal{I})$ follows from (3.2), (3.3), (3.4) and the formula (1.8). □

The MAP partition of $[n]$ with observed $\boldsymbol{x} = (x_i)_{i=1}^n$ is of course not affected by the proportional change of the score formula and hence it is also the maximiser of $\tilde{Q}_{\boldsymbol{x}}(\cdot)$.

## 3.1 Proportional growth of cluster sizes

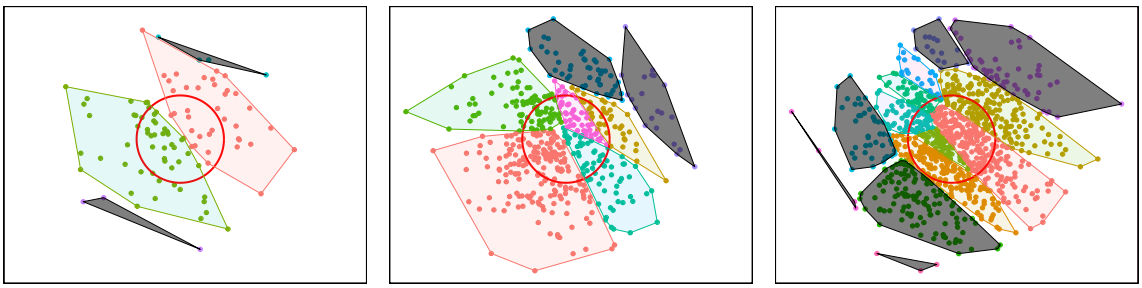
In the sequel, we assume that observations come from iid sampling (which is pursued in Section 3.2 and further). The main goal is to prepare the ground for analysing the limit behaviour of the MAP clustering. To establish this, we want to prove that the sizes of the clusters in the MAP clustering grow proportionally to the quantity of data. Proposition 3.2 gives a partial answer. It 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. The proof of Proposition 3.2 is lengthy and consists of several auxiliary lemmas. It is therefore presented as a separate subsection.

**Proposition 3.2.** *Let* $\sup_n \frac{1}{n}\sum_{i=1}^n \|x_n\|^2 < \infty$ *and let* $\hat{\mathcal{I}}_n$ *be any MAP partition of* $(x_1, \dots, x_n)$ *in Normal-Normal BMM with the Chinese Restaurant prior on the space of partitions (i.e. any partition that maximises* (3.1)*). Then*

$$\liminf_{n\to\infty} \min\{|J| \colon J \in \hat{\mathcal{I}}_n, \exists_{j\in J} \|x_j\| < r\}/n > 0$$

*for every* $r > 0$.



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

Figure 3.1: Illustration of Proposition 3.2 and Corollary 3.3. 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, while ensuring 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; such example is shown in Proposition 3.12.

A simple consequence of Proposition 3.2 is that under these assumptions the number of components in the MAP partition that intersect a given ball is almost surely bounded.

**Corollary 3.3.** *If* $\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2\right)_{n=1}^\infty$ *is bounded then for every* $r > 0$ *the number of clusters that intersect* $B(\mathbf{0}, r)$ *is bounded, i.e.*

$$\limsup_{n\to\infty} |\{J \in \hat{\mathcal{I}}_n \colon \exists_{j\in J} \|x_j\| < r\}| < \infty,$$

*where* $\hat{\mathcal{I}}_n$ *is the same as in Proposition 3.2.*

*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. □

## Proof of Proposition 3.2

For the reader's convenience the proof is split into three parts. In the subsection which we denote 'Preliminary lemmas', we list some facts which are important for our analysis. The subsection 'Important properties of the MAP partition' presents lemmas regarding the MAP, which are further used in the subsection 'Concluding the proof of Proposition 3.2', where the actual proof of the main proposition is finally presented.

### Preliminary lemmas

**Remark 3.4.** Let $R_m$ be defined as in the statement of Remark 3.1, then

(a) $\det R_m$ is a decreasing sequence that converges to $\det R$

(b) if $y_m \to y$ then $R_m y_m \to Ry$

*Proof.* The proof is straightforward and therefore omitted. □

**Lemma 3.5.** *Let $n_1, \ldots, n_k \in \mathbb{N}$, $n_1 \le n_2 \le \ldots \le n_k$ and $n = \sum_{i=1}^k n_i = an_k + r$, where $a \in \mathbb{N}$, $r < n_k$. Then $\prod_{i=1}^k n_i! \le (n_k!)^a n_k(n_k-1)\ldots(n_k-r+1)$ (if $r = 0$, the right-hand side being simply $(n_k!)^a$).*

*Proof.* We prove by induction on $n_k$ that the sequence

$$\mathbf{b} = (\underbrace{1, \ldots, n_k, 1, \ldots, n_k, \ldots, 1, \ldots, n_k}_{a}, n_k - r + 1, n_k - r + 2, \ldots, n_k)$$

may be ordered so that it is term-wise not less than $\mathbf{c} = (1, \ldots, n_1, 1, \ldots, n_2, \ldots, 1, \ldots, n_k)$. Clearly the existence of such ordering establishes the lemma. For $n_k = 1$ this is self evident. For $n_k > 1$ we apply 'greedy' approach. Assume $r > 0$ (the case $r = 0$ follows in a similar way). Put all $n_k$'s from $\mathbf{b}$ in the places of $n_k, n_{k-1}, \ldots, n_{k-a}$ in $\mathbf{c}$. The fact that $n_k \ge n_{k-1} \ge \ldots \ge n_1$ ensures that it is possible and all of $n_k - 1, n_{k-1} - 1, \ldots, n_{k-a} - 1, n_{k-a-1}, \ldots, n_1$ are less or equal to $n_k - 1$. Therefore we may apply inductive assumptions to these numbers thus finishing the proof of the lemma. □

**Lemma 3.6.** *For every $\varepsilon > 0$ there exist $K \in \mathbb{N}$ such that if $n_1, \ldots, n_k \le n/K$, where $n = \sum_{i=1}^k n_i$, then $\sqrt[n]{\prod_{i=1}^k n_i!/n!} < \varepsilon$.*

*Proof.* Assume that $n_1 \le \ldots \le n_k \le n/K$ and let $n = an_k + r$, where $0 \le r < n_k$. By Lemma 3.5 we get that

$$\frac{\prod_{i=1}^k n_i!}{n!} \le \frac{(n_k!)^a (n_k - r + 1)\ldots n_k}{n!} \le \frac{1}{1^{n_k}} \cdot \frac{1}{2^{n_k}} \cdot \ldots \cdot \frac{1}{a^{n_k}} \cdot \frac{1}{(a+1)^r} \le \frac{1}{(a!)^{n_k}}. \tag{3.5}$$

Therefore

$$\sqrt[n]{\frac{\prod_{i=1}^k n_i!}{n!}} \le \frac{1}{\sqrt[\frac{n}{n_k}]{a!}} = \frac{1}{\sqrt[\frac{n}{n_k}]{\lfloor \frac{n}{n_k} \rfloor !}}. \tag{3.6}$$

For $K$ large enough this might be arbitrarily small, so the proof follows. □

### Important properties of the MAP partition

Let us fix a sequence $\boldsymbol{x} = (x_n)_{n=1}^{\infty}$ in $\mathbb{R}^d$ and let $\hat{\mathcal{I}}_n$ by any MAP partition of $(x_1, \dots, x_n)$. In order to facilitate the analysis, we introduce the following notation: let $m_n = \min_{J \in \hat{\mathcal{I}}_n} |J|$ and $M_n = \max_{J \in \hat{\mathcal{I}}_n} |J|$ be the minimum and the maximum cluster size in the partition $\hat{\mathcal{I}}_n$. Moreover for $r > 0$ let

$$m_n^{(r)} = \min\{|J| \colon J \in \hat{\mathcal{I}}_n, \|\overline{\boldsymbol{x}_J}\| < r\}, \quad M_n^{(r)} = \max\{|J| \colon J \in \hat{\mathcal{I}}_n, \|\overline{\boldsymbol{x}_J}\| < r\} \tag{3.7}$$

be the minimal and the maximal cluster size in the partition $\hat{\mathcal{I}}_n$ among the clusters whose center of mass lies in $B(\mathbf{0}, r)$. Finally let

$$m_n^{[r]} = \min\{|J| \colon J \in \hat{\mathcal{I}}_n, \exists_{j \in J} \|x_j\| < r\}, \quad M_n^{[r]} = \max\{|J| \colon J \in \hat{\mathcal{I}}_n, \exists_{j \in J} \|x_j\| < r\} \tag{3.8}$$

be the minimal and the maximal cluster size in the partition $\hat{\mathcal{I}}_n$ among the clusters that intersect the ball $B(\mathbf{0}, r)$.

Let $J_n^m, J_n^M \in \hat{\mathcal{I}}_n$ satisfy $|J_n^m| = m_n$ and $|J_n^M| = M_n$. We choose $J_n^{m,(r)}$, $J_n^{M,(r)}$, $J_n^{m,[r]}$ and $J_n^{M,[r]}$ similarly (e.g. $J_n^{m,(r)} \in \hat{\mathcal{I}}_n$ satisfies $\left\|\overline{\boldsymbol{x}_{J_n^{m,(r)}}}\right\| < r$ and $|J_n^{m,(r)}| = m_n^{(r)}$). Note that this choice may not be unique.

**Proposition 3.7.** *If* $\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2\right)_{n=1}^{\infty}$ *is bounded then* $\liminf_{n\to\infty} M_n/n > 0$.

*Proof.* Suppose that $\liminf M_n/n = 0$. Then there exists an increasing sequence $(n_k)_{k\in\mathbb{N}}$ such that $M_{n_k}/n_k < 1/k$ for every $k \in \mathbb{N}$. We now prove that

$$\lim_{k\to\infty} \sqrt[n_k]{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_{n_k})/\tilde{Q}_{\boldsymbol{x}}(\{[n_k]\})} = 0,$$

hence obtaining a contradiction with the definition of the MAP partition. By (3.1)

$$\begin{aligned}\sqrt[n_k]{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_{n_k})/\tilde{Q}_{\boldsymbol{x}}([n_k])} = \sqrt[n_k]{C^{|\hat{\mathcal{I}}_{n_k}|}/C} \cdot \sqrt[n_k]{\prod_{J\in\hat{\mathcal{I}}_{n_k}} |J|!/n_k!} \cdot \sqrt[n_k]{\frac{{n_k}^{(d+2)/2}\det R_{n_k}}{\prod_{J\in\hat{\mathcal{I}}_{n_k}} |J|^{(d+2)/2}\det R_{|J|}}} \cdot \\ \cdot \exp\left\{\frac{1}{2n_k}\Big(\sum_{J\in\hat{\mathcal{I}}_{n_k}} |J|\left\|R_{|J|}^{-1}R^2\overline{\boldsymbol{x}_J}\right\|^2 - n_k\left\|R_{n_k}^{-1}R^2\overline{x_{[n_k]}}\right\|^2\Big)\right\}.\end{aligned} \tag{3.9}$$

Firstly note that

$$\limsup_{k\to\infty} \sqrt[n_k]{C^{|\hat{\mathcal{I}}_{n_k}|}/C} = \limsup_{k\to\infty} C^{(|\hat{\mathcal{I}}_{n_k}|-1)/n_k} \le \max\{1, C\}. \tag{3.10}$$

By Lemma 3.6, it follows that, under the assumptions,

$$\lim_{k\to\infty} \sqrt[n_k]{\prod_{J\in\hat{\mathcal{I}}_{n_k}} |J|!/n_k!} = 0. \tag{3.11}$$

We have $\sqrt[n_k]{n_k} \to 1$, $R_{n_k} \to R$ and $\det R_{n_k} \geq \det R$. Hence

$$\limsup_{k\to\infty} \sqrt[n_k]{\frac{{n_k}^{(d+2)/2} \det R_{n_k}}{\prod_{J\in\hat{\mathcal{I}}_{n_k}} |J|^{(d+2)/2} \det R_{|J|}}} \leq \frac{\limsup_{k\to\infty} \sqrt[n_k]{{n_k}^{(d+2)/2} \det R_{n_k}}}{\liminf_{k\to\infty} \sqrt[n_k]{\det R^{|\hat{\mathcal{I}}_{n_k}|}}} \leq \frac{1}{\min\{1, \det R\}}. \tag{3.12}$$

Recall the inequality between linear and quadratic means which states that for every sequence $\alpha_1, \ldots, \alpha_l$ of real numbers we have

$$\left|\frac{\sum_{i=1}^l \alpha_i}{l}\right| \leq \sqrt{\frac{\sum_{i=1}^l \alpha_i^2}{l}} \quad \text{or equivalently:} \quad l \cdot \left(\frac{\sum_{i=1}^l \alpha_i}{l}\right)^2 \leq \sum_{i=1}^l \alpha_i^2. \tag{3.13}$$

If we apply (3.13) to every coordinate of vectors $y_1, \ldots, y_d \in \mathbb{R}^d$ and sum up obtained inequalities we obtain that

$$l \cdot \left\|\frac{\sum_{i=1}^l y_i}{l}\right\|^2 \leq \sum_{i=1}^l \|y_i\|^2. \tag{3.14}$$

Therefore, setting $y_i = R_{|J|}^{-1} R^2 x_i$ and using the linearity of multiplication by matrix

$$\sum_{J\in\hat{\mathcal{I}}_n} |J| \left\|R_{|J|}^{-1} R^2 \overline{\boldsymbol{x}_J}\right\|^2 \leq \sum_{J\in\hat{\mathcal{I}}_n} \sum_{j\in J} \left\|R_{|J|}^{-1} R^2 x_j\right\|^2 \tag{3.15}$$

and hence, using Lemma A.6, we have

$$\sum_{J\in\hat{\mathcal{I}}_n} |J| \left\|R_{|J|}^{-1} R^2 \overline{\boldsymbol{x}_J}\right\|^2 \leq \sum_{J\in\hat{\mathcal{I}}_n} \sum_{j\in J} \left\|R_{|J|}^{-1} R^2 x_j\right\|^2 \leq \sum_{J\in\hat{\mathcal{I}}_n} \sum_{j\in J} \left\|R^{-1} R^2 x_j\right\|^2 \leq \|R\|_2^2 \sum_{i\in[n]} \|x_i\|^2, \tag{3.16}$$

where $\|\cdot\|_2$ is a matrix norm induced by $\|\cdot\|$ (i.e. $\|A\|_2 = \sup_{\|x\|=1} \|Ax\|$). From this and assumptions of the Proposition we can easily deduce that

$$\frac{1}{n_k}\Big(\sum_{J\in\hat{\mathcal{I}}_{n_k}} |J| \left\|R_{|J|}^{-1} R^2 \overline{\boldsymbol{x}_J}\right\|^2 - n_k \left\|R_{n_k}^{-1} R^2 \overline{x_{[n_k]}}\right\|^2\Big) \quad \text{is bounded from above.} \tag{3.17}$$

Gathering (3.9), (3.10), (3.11), (3.12) and (3.17) together, we obtain that

$$\limsup_{k\to\infty} \sqrt[n_k]{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_{n_k}) / \tilde{Q}_{\boldsymbol{x}}([n_k])} = 0.$$

Hence there exists a sufficiently large $n$ that satisfies $\mathbb{P}(\hat{\mathcal{I}}_n \,|\, \mathbf{x}) < \mathbb{P}(\{[n]\} \,|\, \mathbf{x})$. This is a contradiction. □

**Corollary 3.8.** *If $\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2\right)_{n=1}^\infty$ is bounded then there exist $r_0 > 0$ such that $\|\overline{x_{J_n^M}}\| \leq r_0$ for all $n > 0$ (and arbitrary choice of $J_n^M$ in case of ambiguity).*

*Proof.* By Proposition 3.7 we know that $\gamma := \liminf_{n\to\infty} M_n/n > 0$, so there exists $N > 0$ such that $M_n/n > \gamma/2$ for $n > N$. Suppose that there exists a sequence $(n_k)_{k=1}^\infty$ such that $\|\overline{\boldsymbol{x}_{J_{n_k}^M}}\| \geq k$ for all $k \in \mathbb{N}$. Note that for $n_k > N$

$$\frac{1}{n_k}\sum_{i=1}^{n_k} \|x_i\| \geq \frac{1}{n_k}\sum_{i\in J_{n_k}^M} \|x_i\| \geq \frac{1}{n_k}\Big\|\sum_{i\in J_{n_k}^M} x_i\Big\| = \frac{M_{n_k}}{n_k}\|\overline{\boldsymbol{x}_{J_{n_k}^M}}\| \geq \gamma/2 \cdot k, \tag{3.18}$$

which, together with the inequality between the arithmetic and quadratic mean, contradicts the assumption that the sequence $\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2\right)_{n=1}^{\infty}$ is bounded. The proof of the Lemma now follows directly. □

**Proposition 3.9.** *If* $\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2\right)_{n=1}^{\infty}$ *is bounded then* $\liminf_{n\to\infty} m_n^{(r)}/n > 0$ *for every* $r > 0$.

*Proof.* Firstly note that it is enough to prove the statement of Proposition 3.9 for all $r > r_0$ for some given $r_0 > 0$ – indeed, for fixed $n \in \mathbb{N}$, $m_n^{(r)}$ is decreasing with $r$. We take $r_0$ from the statement of Corollary 3.8.

Fix $r > r_0$. By the definition of $r_0$ we have $J_n^{M,(r)} = J_n^M$, so $M_n^{(r)} = M_n$ and by Proposition 3.7 $\liminf_{n\to\infty} M_n^{(r)}/n > 0$. Now we prove that $\liminf_{n\to\infty} m_n^{(r)}/n > 0$. Suppose the contrary. We show that for sufficiently large $n$, the posterior probability of $\hat{\mathcal{J}}_n$ increases if we create one cluster out of $J_n^{m,(r)}$ and $J_n^{M,(r)}$. Let $\tilde{\mathcal{I}}_n$ be a partition of $[n]$ obtained from $\hat{\mathcal{I}}_n$ by joining $J_n^{m,(r)}$ with $J_n^{M,(r)}$, i.e.

$$\tilde{\mathcal{I}}_n = \hat{\mathcal{I}}_n \setminus \{J_n^{m,(r)}, J_n^{M,(r)}\} \cup \{J_n^{m,(r)} \cup J_n^{M,(r)}\}. \tag{3.19}$$

In order to simplify the notation, we write $m, M$ instead of $m_n^{(r)}, M_n^{(r)}$ respectively, remembering that they are both functions of $n$. Similarly let us write $\overline{\boldsymbol{x}_m}$, $\overline{\boldsymbol{x}_M}$ and $\overline{\boldsymbol{x}_{m\cup M}}$ instead of $\overline{\boldsymbol{x}_{J_n^{m,(r)}}}, \overline{\boldsymbol{x}_{J_n^{M,(r)}}}$ and $\overline{\boldsymbol{x}_{J_n^{m,(r)}\cup J_n^{M,(r)}}}$. When taking a quotient $\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_n)/\tilde{Q}_{\boldsymbol{x}}(\tilde{\mathcal{I}}_n \,|\, \mathbf{x})$ most factors in (3.1) cancel out, giving

$$\frac{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_n)}{\tilde{Q}_{\boldsymbol{x}}(\tilde{\mathcal{I}}_n)} = C\frac{m!M!}{(m+M)!}\left(\frac{m+M}{mM}\right)^{(d+2)/2}\frac{\det R_{m+M}}{\det R_m \cdot \det R_M}\cdot \exp\{D_n\}^{1/2}, \tag{3.20}$$

where

$$D_n = m\left\|R_m^{-1}R^2\overline{\boldsymbol{x}_m}\right\|^2 + M\left\|R_M^{-1}R^2\overline{\boldsymbol{x}_M}\right\|^2 - (m+M)\left\|R_{m+M}^{-1}R^2\overline{\boldsymbol{x}_{m\cup M}}\right\|^2. \tag{3.21}$$

By Remark 3.4 we have $\det R_{m+M} \le \det R_m$ and $\det R_M \ge R$ and hence

$$\frac{\det R_{m+M}}{\det R_m \cdot \det R_M} \le (\det R)^{-1}. \tag{3.22}$$

Let $I$ be the identity matrix and let $U = \Psi_0^{-1}$ and $V$ be a symmetric matrix such that $V^2 = R^{-1}U^2R^{-1}$. Using Lemma A.7 we get

$$\begin{aligned}(m+M)I - (m+M)R(R_{m+M}^{-1})^2R &= (m+M)\big(I - RR_{m+M}^{-2}R\big) = \\ &= (m+M)\Big(I - \big(I + R^{-1}U^2R^{-1}/(m+M)\big)^{-1}\Big) = \\ &= V\big(I + V^2/(m+M)\big)^{-1}V\end{aligned} \tag{3.23}$$

and therefore

$$\begin{aligned}(m+M)\left\|R\overline{\boldsymbol{x}_{m\cup M}}\right\|^2 - (m+M)\left\|R_{m+M}^{-1}R^2\overline{\boldsymbol{x}_{m\cup M}}\right\|^2 &= \\ = \overline{\boldsymbol{x}_{m\cup M}}^{\top}RV\big(I + V^2/(m+M)\big)^{-1}VR\overline{\boldsymbol{x}_{m\cup M}} &\le \|VR\|_2^2 r^2.\end{aligned} \tag{3.24}$$

Moreover it is straightforward to verify that

$$\left\|R_m^{-1}R^2\overline{\boldsymbol{x}_m}\right\|^2 \le \left\|R\overline{\boldsymbol{x}_m}\right\|^2, \quad \left\|R_M^{-1}R^2\overline{\boldsymbol{x}_M}\right\|^2 \le \left\|R\overline{\boldsymbol{x}_M}\right\|^2, \tag{3.25}$$

and

$$\begin{aligned} m\left\|R\overline{\boldsymbol{x}_m}\right\|^2 + M\left\|R\overline{\boldsymbol{x}_M}\right\|^2 - (m+M)\left\|R\overline{\boldsymbol{x}_{m\cup M}}\right\|^2 = \frac{mM}{m+M}\left\|R(\overline{\boldsymbol{x}_m} - \overline{\boldsymbol{x}_M})\right\|^2 \le \\ \le m\left\|R(\overline{\boldsymbol{x}_m} - \overline{\boldsymbol{x}_M})\right\|^2 \le m\|R\|^2(\|\overline{\boldsymbol{x}_m}\| + \|\overline{\boldsymbol{x}_M}\|)^2 \le m\|R\|^2 \cdot 4r^2, \end{aligned} \tag{3.26}$$

By Lemma A.6, together with (3.24), (3.25) and (3.26),

$$D_n \le m\|R\|_2^2 \cdot 4r^2 + \|VR\|_2^2 r^2. \tag{3.27}$$

Stirling formula, which is valid for every $n \in \mathbb{N}$ (cf. Feller (1968)), states that

$$\sqrt{2\pi n}(n/e)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}(n/e)^n e^{\frac{1}{12n}}. \tag{3.28}$$

This gives:

$$\frac{m!M!}{(m+M)!} \le \sqrt{2\pi}\left(\frac{mM}{m+M}\right)^{1/2}\frac{m^m M^M}{(m+M)^{(m+M)}}e \le \sqrt{2\pi}e\left(\frac{mM}{m+M}\right)^{1/2}\left(\frac{m}{M}\right)^m. \tag{3.29}$$

Now by applying (3.22), (3.27) and (3.29) to (3.20) we obtain that for constants $C'$ and $C''$, given by

$$C' = C\sqrt{2\pi}(\det R)^{-1}\exp\{\|VR\|_2^2 r^2\}, \quad C'' = \exp\{\|R\|_2^2 \cdot 4r^2\}, \tag{3.30}$$

such that

$$\liminf_{n\to\infty}\frac{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_n)}{\tilde{Q}_{\boldsymbol{x}}(\tilde{\mathcal{I}}_n)} \le \liminf_{n\to\infty} C'\left(\frac{m+M}{mM}\right)^{(d+1)/2}\left(\frac{mC''}{M}\right)^m = 0, \tag{3.31}$$

as $\liminf_{n\to\infty} m/M \to 0$. Hence there exist $n$ such that the posterior probability of $\hat{\mathcal{I}}_n$ is smaller than the posterior probability of $\tilde{\mathcal{I}}_n$. This contradicts the definition of $\hat{\mathcal{I}}_n$ and finishes the proof of the Lemma. □

## Concluding the proof of Proposition 3.2

Assume that $\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2\right)_{n=1}^\infty$ is bounded. We want to prove that $\liminf_{n\to\infty} m_n^{[r]}/n > 0$ for every $r > 0$.

Take $r_0$ from the statement of Corollary 3.8. Note that, as in proof of Proposition 3.9 it is enough to prove the statement of Proposition 3.2 for $r > r_0$.

Fix $r > r_0$. Suppose that $\liminf_{n\to\infty} m_n^{[r]}/n = 0$ and let $(n_k)_{k=1}^\infty$ be a sequence such that $\lim_{k\to\infty} m_{n_k}^{[r]}/n_k = 0$. This implies that

$$\lim_{k\to\infty}\left\|\overline{\boldsymbol{x}_{J_{n_k}^{m,[r]}}}\right\| = \infty \tag{3.32}$$

(otherwise we would obtain a contradiction with Proposition 3.9). Let

$$I_n^a = \{j \in J_n^{m,[r]} \colon \|x_j\| \le r\}, \quad I_n^b = \{j \in J_n^{m,[r]} \colon \|x_j\| > r\}. \tag{3.33}$$

Consider a partition $\hat{\mathcal{J}}_n$ obtained from $\hat{\mathcal{I}}_n$ by taking $I_n^a$ from $J_n^{m,[r]}$ and adding it to $J_n^M$, i.e.

$$\hat{\mathcal{J}}_n = \hat{\mathcal{I}}_n \setminus \{J_n^{m,[r]}, J_n^M\} \cup \{J_n^{m,[r]} \setminus I_n^a, J_n^M \cup I_n^a\}. \tag{3.34}$$

When taking a quotient $\frac{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_n)}{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{J}}_n)}$ most factors in (3.1) cancel out, giving

$$\frac{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_n)}{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{J}}_n)} = \frac{(a+b)!M!}{b!(a+M)!}\left(\frac{b(a+M)}{(a+b)M}\right)^{(d+2)/2} \frac{\det R_b \cdot \det R_{a+M}}{\det R_{a+b} \cdot \det R_M} \cdot \exp\left\{\frac{1}{2}\check{D}_{n_k}\right\}, \tag{3.35}$$

where $M = |J_{n_k}^M|$, $a = |I_{n_k}^a|$, $b = |I_{n_k}^b|$ (in order to simplify the notation we skip the index $n_k$) and

$$\begin{aligned}\check{D}_{n_k} = (a+b)\left\|R_{a+b}^{-1}R^2\overline{\boldsymbol{x}_{a\cup b}}\right\|^2 + M\left\|R_M^{-1}R^2\overline{\boldsymbol{x}_M}\right\|^2 - \\ - b\left\|R_b^{-1}R^2\overline{\boldsymbol{x}_b}\right\|^2 - (a+M)\left\|R_{a+M}^{-1}R^2\overline{\boldsymbol{x}_{a\cup M}}\right\|^2.\end{aligned} \tag{3.36}$$

in which $\overline{\boldsymbol{x}_{a\cup b}} = \overline{\boldsymbol{x}_{I_{n_k}^a \cup I_{n_k}^b}}$ and we define $\overline{\boldsymbol{x}_b}, \overline{\boldsymbol{x}_M}, \overline{\boldsymbol{x}_{a\cup M}}$ similarly. Note that

$$\frac{(a+b)!M!}{b!(a+M)!} = \frac{(b+1)^{(a)}}{(M+1)^{(a)}} < \frac{b+1}{M+1} \overset{k\to\infty}{\longrightarrow} 0, \tag{3.37}$$

since $\lim_{k\to\infty}(a+b)/n_k = \lim_{k\to\infty} m_{n_k}^{[r]}/n_k = 0$ and $\liminf_n M/n > 0$. For the similar reason

$$\frac{b(a+M)}{(a+b)M} < \frac{a+M}{M} \overset{k\to\infty}{\longrightarrow} 1. \tag{3.38}$$

Moreover, by Remark 3.4 we have

$$\frac{\det R_b \cdot \det R_{a+M}}{\det R_{a+b} \cdot \det R_M} \le \frac{\det R_1^2}{\det R^2}. \tag{3.39}$$

Now let us investigate $\check{D}_n$. The notation is easier after a linear substitution $y_i = R^2 x_i$ (so that $\overline{\mathbf{y}_I} = R^2\overline{\boldsymbol{x}_I}$), hence obtaining

$$\check{D}_{n_k} = (a+b)\left\|R_{a+b}^{-1}\overline{\mathbf{y}_{a\cup b}}\right\|^2 + M\left\|R_M^{-1}\overline{\mathbf{y}_M}\right\|^2 - b\left\|R_b^{-1}\overline{\mathbf{y}_b}\right\|^2 - (a+M)\left\|R_{a+M}^{-1}\overline{\mathbf{y}_{a\cup M}}\right\|^2. \tag{3.40}$$

Note that

$$\begin{aligned}(a+b)\left\|R_{a+b}^{-1}\overline{\mathbf{y}_{a+b}}\right\|^2 - b\left\|R_b^{-1}\overline{\mathbf{y}_b}\right\|^2 &= (a+b)\left\|R_{a+b}^{-1}\left(\frac{a}{a+b}\overline{\mathbf{y}_a} + \frac{b}{a+b}\overline{\mathbf{y}_b}\right)\right\|^2 - b\left\|R_b^{-1}\overline{\mathbf{y}_b}\right\|^2 = \\ &= \frac{1}{a+b}\left\|R_{a+b}^{-1}(a\overline{\mathbf{y}_a} + b\overline{\mathbf{y}_b})\right\|^2 - b\left\|R_b^{-1}\overline{\mathbf{y}_b}\right\|^2 \le \\ &= \frac{1}{a+b}\Big(a^2\left\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\right\|^2 + 2ab\left\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\right\|\left\|R_{a+b}^{-1}\overline{\mathbf{y}_b}\right\| + b^2\left\|R_{a+b}^{-1}\overline{\mathbf{y}_b}\right\|^2 - \\ &\quad - b(a+b)\left\|R_b^{-1}\overline{\mathbf{y}_b}\right\|^2\Big) = \\ &= \frac{1}{a+b}\Big(a^2\left\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\right\|^2 + 2ab\left\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\right\|\left\|R_{a+b}^{-1}\overline{\mathbf{y}_b}\right\| - \overline{\mathbf{y}_b}T_1\overline{\mathbf{y}_b}\Big)\end{aligned} \tag{3.41}$$

where $T_1 = b(a+b)R_b^{-2} - b^2R_{a+b}^{-2}$. For two positive definite matrices $M_1, M_2$ we write $M_1 \succeq M_2$ when $M_1 - M_2$ is positive definite. If $A, B$ are two invertible matrices then

$$A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1} \tag{3.42}$$

and hence

$$\begin{aligned}
T_{a,b} &= b(a+b)(R^2+U^2/b)^{-1} - b^2\big(R^2+U^2/(a+b)\big)^{-1} = \\
&= b^2(a+b)(bR^2+U^2)^{-1} - b^2(a+b)\big((a+b)R^2+U^2\big)^{-1} = \\
&= b^2(a+b)\Big((bR^2+U^2)^{-1} - \big((a+b)R^2+U^2\big)^{-1}\Big) = \\
&= ab^2(a+b)\big((a+b)R^2+U^2\big)^{-1}R^2(bR^2+U^2)^{-1} = \\
&= ab\big(R^2+U^2/(a+b)\big)^{-1}R^2(R^2+U^2/b)^{-1}.
\end{aligned} \tag{3.43}$$

Note that by a direct calculation:

$$(R^2+U^2/b)R^{-2}\big(R^2+U^2/(a+b)\big) \preceq (R^2+U^2)R^{-2}\big(R^2+U^2\big) \tag{3.44}$$

and hence, by Lemma A.6 (d) and (3.43)

$$T_{a,b} \succeq ab(R^2+U^2)^{-1}R^2(R^2+U^2)^{-1}. \tag{3.45}$$

Let $T_0$ be a symmetric, positive definite matrix such that $T_0^2 = (R^2+U^2)^{-1}R^2(R^2+U^2)^{-1}$. Using (3.41) and (3.45) we have that

$$\begin{aligned}
(a+b)\big\|R_{a+b}^{-1}\overline{\mathbf{y}_{a+b}}\big\|^2 - b\big\|R_b^{-1}\overline{\mathbf{y}_b}\big\|^2 &\le \frac{1}{a+b}\big(a^2\big\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\big\|^2 + 2ab\big\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\big\|\big\|R_{a+b}^{-1}\overline{\mathbf{y}_b}\big\| - ab\big\|T_0\overline{\mathbf{y}_b}\big\|^2\big) = \\
&= a\Big(\frac{a}{a+b}\big\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\big\|^2 + \frac{b}{a+b}\big(2\big\|R_{a+b}^{-1}\overline{\mathbf{y}_a}\big\|\big\|R_{a+b}^{-1}\overline{\mathbf{y}_b}\big\| - \big\|T_0\overline{\mathbf{y}_b}\big\|^2\big)\Big) \le \\
&\le a\Big(\frac{a}{a+b}\big\|R^{-1}\overline{\mathbf{y}_a}\big\|^2 + \frac{b}{a+b}\big(2\big\|R^{-1}\overline{\mathbf{y}_a}\big\|\big\|R^{-1}\overline{\mathbf{y}_b}\big\| - \big\|T_0\overline{\mathbf{y}_b}\big\|^2\big)\Big).
\end{aligned} \tag{3.46}$$

Let $\underline{\nu}_A$ be the minimal eigenvalue of the square matrix $A$. Then for any symmetric, positive definite matrix $A$ and vector $v$ we have $\underline{\nu}_A\|v\| \le \|Av\| \le \overline{\nu}_A\|v\|$ and $\overline{\nu}_A = \underline{\nu}_{A^{-1}}^{-1}$. Hence, by (3.46)

$$(a+b)\big\|R_{a+b}^{-1}\overline{\mathbf{y}_{a+b}}\big\|^2 - b\big\|R_b^{-1}\overline{\mathbf{y}_b}\big\|^2 \le a\Big(\underline{\nu}_R^{-2}\frac{a}{a+b}\big\|\overline{\mathbf{y}_a}\big\|^2 + \frac{b}{a+b}\big\|\overline{\mathbf{y}_b}\big\|\big(2\underline{\nu}_R^{-2}\big\|\overline{\mathbf{y}_a}\big\| - \underline{\nu}_{T_0}^2\big\|\overline{\mathbf{y}_b}\big\|\big)\Big) \tag{3.47}$$

Similarly we note that

$$\begin{aligned}
&M\big\|R_M^{-1}\overline{\mathbf{y}_M}\big\|^2 - (a+M)\big\|R_{a+M}^{-1}\overline{\mathbf{y}_{a\cup M}}\big\|^2 \le \\
&\quad = M\big\|R_M^{-1}\overline{\mathbf{y}_M}\big\|^2 - \frac{1}{a+M}\big(a^2\big\|R_{a+M}^{-1}\overline{\mathbf{y}_a}\big\|^2 + 2aM\overline{\mathbf{y}_a}^\top R_{a+M}^{-2}\overline{\mathbf{y}_M} + M^2\big\|R_{a+M}^{-1}\overline{\mathbf{y}_M}\big\|^2\big) = \\
&\quad = \frac{1}{a+M}\big((a+M)M\big\|R_M^{-1}\overline{\mathbf{y}_M}\big\|^2 - a^2\big\|R_{a+M}^{-1}\overline{\mathbf{y}_a}\big\|^2 - 2aM\overline{\mathbf{y}_a}^\top R_{a+M}^{-2}\overline{\mathbf{y}_M} - M^2\big\|R_{a+M}^{-1}\overline{\mathbf{y}_M}\big\|^2\big) \le \\
&\quad \le \frac{1}{a+M}\Big(M\big((a+M)\big\|R_M^{-1}\overline{\mathbf{y}_M}\big\|^2 - M\big\|R_{a+M}^{-1}\overline{\mathbf{y}_M}\big\|^2\big) - 2aM\overline{\mathbf{y}_a}^\top R_{a+M}^{-2}\overline{\mathbf{y}_M}\Big).
\end{aligned} \tag{3.48}$$

Using (3.42) again, we can write

$$\begin{aligned}
(a+M)R_M^{-2} - MR_{a+M}^{-2} &= (a+M)(R^2+U^2/M)^{-1} - M\big(R^2+U^2/(a+M)\big)^{-1} = \\
&= M(a+M)(MR^2+U^2)^{-1} - M(a+M)\big((a+M)R^2+U^2\big)^{-1} = \\
&= M(a+M)\Big((MR^2+U^2)^{-1} - \big((a+M)R^2+U^2\big)^{-1}\Big) = \\
&= aM(a+M)\big((a+M)R^2+U^2\big)^{-1}R^2(MR^2+U^2)^{-1} = \\
&= a\big(R^2+U^2/(a+M)\big)^{-1}R^2(R^2+U^2/M)^{-1}.
\end{aligned} \tag{3.49}$$

By direct calculation

$$(R^2 + U^2/M)R^{-2}\big(R^2 + U^2/(a+M)\big) \succeq R^2 \tag{3.50}$$

and hence, by Lemma A.6 (d) and (3.49) we get

$$(a+M)R_M^{-2} - MR_{a+M}^{-2} \preceq aR^{-2}. \tag{3.51}$$

By (3.48) and (3.51)

$$\begin{aligned} M\big\|R_M^{-1}\overline{\mathbf{y}_M}\big\|^2 - (a+M)\big\|R_{a+M}^{-1}\overline{\mathbf{y}_{a\cup M}}\big\|^2 &\le \frac{1}{a+M}\big(aM\big\|R^{-1}\overline{\mathbf{y}_M}\big\|^2 - 2aM\overline{\mathbf{y}_a}^\top R_{a+M}^{-2}\overline{\mathbf{y}_M}\big) = \\ &= a\frac{M}{a+M}\big(\big\|R^{-1}\overline{\mathbf{y}_M}\big\|^2 - 2\overline{\mathbf{y}_a}^\top R_{a+M}^{-2}\overline{\mathbf{y}_M}\big) \le \\ &\le a\frac{M}{a+M}\big(\big\|R^{-1}\overline{\mathbf{y}_M}\big\|^2 + 2\big\|R_{a+M}^{-1}\overline{\mathbf{y}_a}\big\|\cdot\big\|R_{a+M}^{-1}\overline{\mathbf{y}_M}\big\|\big) \le \\ &\le a\frac{M}{a+M}\big(\big\|R^{-1}\overline{\mathbf{y}_M}\big\|^2 + 2\big\|R^{-1}\overline{\mathbf{y}_a}\big\|\cdot\big\|R^{-1}\overline{\mathbf{y}_M}\big\|\big). \end{aligned} \tag{3.52}$$

Joining (3.47) and (3.52) we get that

$$\begin{aligned} \check{D}_{n_k} \le a\Big(&\underline{\nu}_R^{-2}\frac{a}{a+b}\big\|\overline{\mathbf{y}_a}\big\|^2 + \frac{b}{a+b}\big\|\overline{\mathbf{y}_b}\big\|\big(2\underline{\nu}_R^{-2}\big\|\overline{\mathbf{y}_a}\big\| - \underline{\nu}_{T_2}^2\big\|\overline{\mathbf{y}_b}\big\|\big) + \\ &+ \frac{M}{a+M}\big(\big\|R\overline{\mathbf{y}_M}\big\|^2 + 2\big\|R^{-1}\overline{\mathbf{y}_a}\big\|\cdot\big\|R^{-1}\overline{\mathbf{y}_M}\big\|\big)\Big) \end{aligned} \tag{3.53}$$

By the triangle inequality

$$\begin{aligned} \frac{b}{a+b}\big\|\overline{\mathbf{y}_b}\big\| &\ge \big\|\overline{\mathbf{y}_{a\cup b}}\big\| - \frac{a}{a+b}\big\|\overline{\mathbf{y}_a}\big\| \ge \underline{\nu}_R^2\big\|\overline{\boldsymbol{x}_{a\cup b}}\big\| - \frac{a}{a+b}\overline{\nu}_R^2\big\|\overline{\mathbf{y}_a}\big\| \ge \\ &\ge \underline{\nu}_R^2\big\|\overline{\boldsymbol{x}_{a\cup b}}\big\| - \frac{a}{a+b}\overline{\nu}_R^2 r^2. \end{aligned} \tag{3.54}$$

Hence, by (3.32)

$$\lim_{k\to\infty}\frac{b}{a+b}\big\|\overline{\mathbf{y}_b}\big\| = \infty. \tag{3.55}$$

Note that in particular $\lim_{k\to\infty}\big\|\overline{\mathbf{y}_b}\big\| = \infty$ and since $\big\|\overline{\mathbf{y}_a}\big\| \le \overline{\nu}_{R^2}\big\|\overline{\boldsymbol{x}_a}\big\| \le \overline{\nu}_{R^2}r$ we have

$$\lim_{k\to\infty}\big(2\underline{\nu}_R^{-2}\big\|\overline{\mathbf{y}_a}\big\| - \underline{\nu}_{T_2}^2\big\|\overline{\mathbf{y}_b}\big\|\big) = -\infty. \tag{3.56}$$

Moreover $\big\|\overline{\mathbf{y}_M}\big\| \le \overline{\nu}_{R^2}\big\|\overline{\boldsymbol{x}_M}\big\| \le \overline{\nu}_{R^2}r$ and therefore by (3.53), (3.54), (3.39) and (3.56) we have

$$\lim_{k\to\infty} D_{n_k} = -\infty \tag{3.57}$$

By taking (3.35) and using (3.37), (3.38) and (3.57) we obtain that $\lim_{k\to\infty}\frac{\tilde{Q}_{\boldsymbol{x}}(\hat{\mathcal{I}}_{n_k})}{\bar{Q}_{\boldsymbol{x}}(\hat{\mathcal{J}}_{n_k})} = 0$. This is a contradiction, from which the result follows.

## 3.2 Clustering 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 $P$. Now let $\hat{\mathcal{I}}_n$ by any MAP partition of $(X_1, \ldots, X_n)$ (i.e. any maximiser of (3.1); in case of ambiguity just choose one at random). Then $\hat{\mathcal{I}}_n$ goes a random partition of $[n]$. Note that if $E_P\|X\|^4 < \infty$ then by the Strong Law of Large Numbers almost surely $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \to E_P\|X\|^2 < \infty$ and therefore the assumptions of Proposition 3.2 are satisfied almost surely. Useful corollaries of this observation are listed below.

**Corollary 3.10.** *If $E_P\|X\|^4 < \infty$ then almost surely for every $r > 0$*

*(a)* $\liminf_{n\to\infty} \min\{|J| \colon J \in \hat{\mathcal{I}}_n, \exists_{j\in J}\|X_j\| < r\}/n > 0$.

*(b) the number of clusters in $\hat{\mathcal{I}}_n$ that intersect $B(\mathbf{0}, r)$ is bounded.*

An easy consequence of Corollary 3.10 is

**Corollary 3.11.** *If the support of $P$ is bounded then almost surely*

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

*(b) $|\hat{\mathcal{I}}_n|$ is bounded.*

*Proof.* If the support of $P$ is bounded then $\mathbb{E}\,\|X^4\| < \infty$. Therefore we can use Corollary 3.10 where we take $r$ sufficiently large so that $B(\mathbf{0}, r)$ contains the support of $P$. □

The assumptions of Corollary 3.11 cannot be relaxed to those of Corollary 3.10. 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 $\{x_{n_k}\} \in \hat{\mathcal{I}}_{n_k}$. This violates part (a) of Corollary 3.11.

**Proposition 3.12.** *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 $\liminf_{n\to\infty} m(\hat{\mathcal{I}}_n) = 1$.*

*Proof.* Take $d = 1$ and $\alpha = \Psi_0 = \Sigma_0 = 1$. Then $R_k = \sqrt{\frac{k+1}{k}}$. Let $y_1, \ldots, y_n \in \mathbb{R}^d$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Take any partition $\mathcal{J}$ of $[n]$. Let $J_n \in \mathcal{J}$ be the cluster containing $n$ and assume that $|J_n| \geq 2$. Let $\mathcal{J}_{n,\{n\}}$ be obtained by creating a singleton out of $n$, i.e. $\mathcal{J}_{n,\{n\}} = \mathcal{J} \setminus \{J_n\} \cup \{J_n \setminus \{n\}, \{n\}\}$. By (3.1) it is easy to show that the quotient $\tilde{Q}_{\boldsymbol{y}}(\mathcal{J}_{n,\{n\}})/\tilde{Q}_{\boldsymbol{y}}(\mathcal{J})$ is equal to

$$h_{J_n}(y_1, \ldots, y_n) = \frac{1}{|J_n| - 1}\sqrt{\frac{|J_n| + 1}{2|J_n|}} \exp\left\{\frac{y_n^2}{4} + \frac{(\sum \boldsymbol{y}_{J_n\setminus\{n\}})^2}{2|J_n|} - \frac{(\sum \boldsymbol{y}_{J_n})^2}{2(|J_n| + 1)}\right\}. \tag{3.58}$$

The exponent in the formula above is equal to

$$y_n^2 \frac{|J_n| - 1}{4(|J_n| + 1)} - y_n \frac{\sum \boldsymbol{y}_{J_n\setminus\{n\}}}{|J_n| + 1} + \frac{(\sum \boldsymbol{y}_{J_n\setminus\{n\}})^2}{2|J_n|(|J_n| + 1)}, \tag{3.59}$$

which is a convex quadratic function of $y_n$. Now, since $|J_n| \geq 2$, it follows that

$$\frac{|J_n|-1}{4(|J_n|+1)} \geq \frac{1}{12} \quad \text{and} \quad \Big|\frac{\sum \boldsymbol{y}_{J_n\setminus\{n\}}}{|J_n|+1}\Big| \leq |\overline{\boldsymbol{y}_{J_n\setminus\{n\}}}|. \tag{3.60}$$

Now let $L = 2\cdot 18^4$ and $\tilde{x}_m = 18^m$. We show that if

$$n \leq L^{m+1}, \quad y_n \geq \tilde{x}_m, \quad \text{and} \quad |y_1|,\dots,|y_{n-1}| \leq \tilde{x}_{m-1} \tag{$\star$}$$

then $h_{J_n}(y_1,\dots,y_n) > 1$ (regardless of $J_n$) and hence in MAP partition for $[n]$ based on data $(y_i)_{i=1}^n$ singleton $\{n\}$ forms a separate cluster. Assume ($\star$). Note that if $n \leq L^{m+1}$ and $|y_1|,\dots,|y_{n-1}| \leq \tilde{x}_{m-1}$ then by (3.58), (3.59) and (3.60) we obtain that

$$h_{J_n}(y_1,\dots,y_n) \geq \frac{1}{L^{m+1}}\sqrt{\frac{1}{2}}\exp\left\{\frac{1}{12}y_n^2 - \tilde{x}_{m-1}y_n\right\}. \tag{3.61}$$

Let us denote the right-hand side of (3.61) by $l(y_n)$. Now $l(y_n) \geq 1$ is equivalent to

$$\frac{1}{12}y_n^2 - \tilde{x}_{m-1}y_n - \big((m+1)\log L + (\log 2)/2\big) \geq 0. \tag{3.62}$$

By the properties of the quadratic function, (3.62) is implied by

$$y_n \geq 6\Big(\tilde{x}_{m-1} + \sqrt{\tilde{x}_{m-1}^2 + \frac{1}{3}\big((m+1)\log L + (\log 2)/2\big)}\,\Big). \tag{3.63}$$

It can be easily proved by induction that $3\tilde{x}_{m-1}^2 > \frac{1}{3}\big((m+1)\log L + (\log 2)/2\big)$ for $m \geq 2$ (note that the left-hand side is geometric with respect to $m$, while the right-hand side is linear) and therefore

$$6\Big(\tilde{x}_{m-1} + \sqrt{\tilde{x}_{m-1}^2 + \frac{1}{3}\big((m+1)\log L + (\log 2)/2\big)}\,\Big) < 18\tilde{x}_{m-1} = \tilde{x}_m \tag{3.64}$$

and as $y_n \geq \tilde{x}_m$ we have that $h_{J_n}(y_1,\dots,y_n) > 1$.

Note that if $(y_n)_{n=1}^\infty$ is a sequence whose terms belong to $\{\tilde{x}_m \colon m \in \mathbb{N}\}$ then if for some $m \in \mathbb{N}$

$$n \leq L^{m+1}, \quad y_n \geq \tilde{x}_m, \quad \text{and} \quad y_1,\dots,y_{n-1} < y_n \tag{$\star'$}$$

then condition ($\star$) holds with some $m' \geq m$ (the one that satisfies $\tilde{x}_{m'} = y_n$). Indeed, if ($\star'$) is satisfied and $y_n = \tilde{x}_{m'}$ then as $y_1,\dots,y_{n-1} < y_n$ we have $y_1,\dots,y_{n-1} \leq \tilde{x}_{m'-1}$, moreover $\tilde{x}_{m'} = y_n \geq \tilde{x}_m$ and hence $m' \geq m$ and $n \leq L^{m+1} \leq L^{m'+1}$ and hence ($\star$) is satisfied.

We now give an example of probability weights $(p_m)_{m\geq 1}$ such that the following probability distribution $P = \sum_{m=1}^\infty p_m\delta_{\tilde{x}_m}$ has a finite fourth moment and if $(X_n)_{n=1}^\infty \overset{\text{iid}}{\sim} P$ then ($\star'$) happens almost surely infinitely many times. Let $q = L^{-1}$ and $p_m = (1-q)q^{m-1}$. It is straightforward to check that in this case $P$ has finite fourth moment, as

$$\sum_{m=1}^\infty p_m\tilde{x}_m^4 = (1-L^{-1})\sum_{m=1}^\infty \frac{(18^m)^4}{(2\cdot 18^4)^{m-1}} = 18^4(1-L^{-1})\sum_{m=1}^\infty \frac{1}{2^{m-1}} < \infty. \tag{3.65}$$

Now let $s_m = \sum_{i=1}^m p_i = 1 - q^m$. Then $s_m^{L^m} = (1 - L^{-m})^{L^m} \to e^{-1}$. Let

$$n_m = \sum_{i=0}^{m} L^i = \frac{L^{m+1} - 1}{L - 1} < L^{m+1} \tag{3.66}$$

and let $A_m$ be an event defined by

$$A_m = \{\max_{n_{m-1} \le i < n_m} X_i \ge \tilde{x}_m\} = \bigcup_{n_{m-1} \le i < n_m} \{X_i \ge \tilde{x}_m\}. \tag{3.67}$$

Then the probability of $A_m$ is equal to $1 - s_{m-1}^{L^m}$ which converges to $1 - e^{-L}$. By the Borel-Cantelli Lemma, it follows that almost surely infinitely many of the events $A_m$ happens. Let $(x_n)_{n=1}^\infty$ be a realisation of $(X_n)_{n=1}^\infty$ and let $(m_k)_{k=1}^\infty$ be an increasing sequence of all indices $m$ for which $A_m$ hold. Now let

$$\hat{n}_m = \min\{n_{m-1} \le n < n_m \colon x_n = \max_{n_{m-1} \le i < n_m} x_i\}. \tag{3.68}$$

Let $(k_i)_{i=1}^\infty$ be an increasing sequence of indices such that $x_{\hat{n}_k} < x_{\hat{n}_{k_i}}$ for $k < k_i$ (such sequence exists since $\tilde{x}_m \to \infty$). We now show that for every $i \in \mathbb{N}$ the condition ($\star'$) is satisfied with $n = \hat{n}_{m_{k_i}}$ and $m = m_{k_i}$. Note that, by the definition, $\hat{n}_m < n_m$ and hence, using (3.66), $\hat{n}_{m_{k_i}} < n_{m_{k_i}} < L^{m_{k_i}+1}$. By the definition of $m_k$, $x_{\hat{n}_{m_k}} \ge \tilde{x}_{m_k}$ for $k \in \mathbb{N}$, and hence $x_{\hat{n}_{m_{k_i}}} \ge \tilde{x}_{m_{k_i}}$. Finally, setting $m(l) = \min\{m \in \mathbb{N} \colon n_m > l\}$, we have

$$\text{for } l < \hat{n}_{m_{k_i}} \text{ we have } \begin{cases} x_l < x_{\hat{n}_{m_{k_i}}} & \text{if } m(l) = m_{k_i}, \\ x_l \le x_{\hat{n}_{m(l)}} < x_{\hat{n}_{m_{k_i}}} & \text{if } m(l) = m_k \text{ for some } k < k_i, \\ x_l < \tilde{x}_{m(l)} < \tilde{x}_{m_{k_i}} \le x_{\hat{n}_{m_{k_i}}} & \text{otherwise,} \end{cases} \tag{3.69}$$

This proves that almost surely the MAP partition creates a new cluster out of a new observation infinitely many times. □

## 3.3 Convergence of the MAP partitions

Corollary 2.13 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 2.2, as the sample size goes to infinity, and under what conditions. This section presents partial answers in this regard, concerning the Normal-Normal BMM with the Chinese Restaurant Process prior on the space of clusterings. Recall that in this case Corollary 2.7 guarantees that the clusters in the MAP clustering are linearly separated. In other words, the *convex hulls* of the clusters in the MAP partition are disjoint. Moreover in this model we already know what is the asymptotic limit of the logarithm of the posterior probability, which for an easier reference and correspondence to our main formula (3.1) is given below.

**Lemma 3.13.** *Let $\mathcal{A}$ be a finite $P$-partition of $\mathbb{R}^d$ consisting of Borel sets with positive $P$ measure. Let $X_1, \ldots \overset{iid}{\sim} P$. Then almost surely*

$$\sqrt[n]{Q_{\mathbf{X}_{1:n}}\left(\mathcal{I}_n^{\mathcal{A}}(\mathbf{X}_{1:n})\right)} \simeq \frac{n}{e} \exp\left\{\Delta_P^{NN}(\mathcal{A})\right\}, \tag{3.70}$$

*where $\Delta_P^{NN}(\mathcal{A})$ is given by (2.60).*

*Proof.* Relation (3.70) could be established by an examination of Theorem 2.12 and the relation between $Q_{\boldsymbol{x}}(\mathcal{I})$ and $Q(\boldsymbol{x}, \mathcal{I})$. For the sake of the remaining part of this chapter, we choose a more straightforward approach a compute the limit directly.

We abuse the notation slightly and denote $p_{I_n^A} = |I_n^A|/n$ for $A \in \mathcal{A}$. By the Strong Law of Large Numbers the sequence $(X_n)_{n=1}^{\infty}$ almost surely satisfies $p_{I_n^A} \to p_A > 0$. By Stirling formula

$$\prod_{I \in \mathcal{I}_n^A} (np_I)! \simeq \prod_{I \in \mathcal{I}_n^A} \left[ \left(\frac{np_I}{e}\right)^{np_I} \sqrt{2\pi np_I} \right] = \sqrt{2\pi n}^{|\mathcal{I}_n^A|} \sqrt{\prod_{I \in \mathcal{I}_n^A} p_I} \cdot \left(\frac{n}{e} \prod_{I \in \mathcal{I}_n^A} p_I^{p_I}\right)^n \tag{3.71}$$

from which it follows by the Strong Law of Large Numbers that $\sqrt[n]{\prod_{I \in \mathcal{I}_n^A} (np_I)!} \simeq \frac{n}{e} \prod_{I \in \mathcal{I}_n^A} p_I^{p_I} \simeq \frac{n}{e} \prod_{I \in \mathcal{A}} p_A^{p_A}$. Note that since $\mathcal{I}_n^{\mathcal{A}}$ has at most $|\mathcal{A}|$ elements,

$$\lim_{n \to \infty} \sqrt[n]{C^{|\mathcal{I}_n^{\mathcal{A}}|}} = 1 \quad \text{and} \quad \lim_{n \to \infty} \sqrt[n]{\prod_{I \in \mathcal{I}_n^{\mathcal{A}}} |I|^{(d+2)/2} \det R_{|I|}} = 1. \tag{3.72}$$

It follows from the Strong Law of Large Numbers that $\overline{\mathbf{X}_{I_n^A}} \to \mathbb{E}\left(X \mid X \in A\right)$ for $A \in \mathcal{A}$ almost surely. It follows that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{I \in \mathcal{I}_n^{\mathcal{A}}} |I| \left\| R_{|I|}^{-1} R^2 \overline{\mathbf{X}_I} \right\|^2 = \sum_{A \in \mathcal{A}} p_A \left\| R\mathbb{E}\left(X \mid X \in A\right) \right\|^2. \tag{3.73}$$

Applying (3.71), (3.72) and (3.73) together with (2.60) to the formula (3.1) for $\mathcal{I}^{\mathcal{A}}$ completes the proof of the Lemma. □

Since the clusters in the MAP partition are convex sets, it seems promising to try to analyse the posterior score $Q$ of the MAP partition in the way similar to the analysis of the induced partition. In order to do so, we would like to have a form of 'uniform law of large numbers' with respect to the family of convex sets. More precisely if $P$ is a probability distribution on $\mathbb{R}^d$ and $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$, for $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ 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 3.14.** *If for each convex set $C$ the boundary $\partial C$ 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 3.14 are satisfied if $P$ has a density with respect to Lebesgue measure $\lambda_d$ on $\mathbb{R}^d$ (since in this case the Lebesgue measure $\lambda_d$ of the boundary of any convex set is 0, and hence is also $P$ measure 0).

Apart from uniformly approximating the probabilities of convex sets with their empirical counterparts, we will also need to do a similar estimation for the conditional expected values. This can be done, provided we take into account convex sets whose probability is separated from 0. This is stated as Lemma 3.15, whose proof is left for the Appendix.

**Lemma 3.15.** *If $P$ satisfies $(*)$ and for $X \sim P$ we have $E_P\|X\|^2 < \infty$ then for every $\delta > 0$ we have*

$$\lim_{n\to\infty} \sup_{\substack{C \text{ convex} \\ P(C)>\delta}} \big\|E_n(X \mid X \in C) - E_P(X \mid X \in C)\big\| = 0 \quad \textit{almost surely.} \tag{3.74}$$

We can now formulate a functional relation between the posterior probability of the MAP partition and the value of the function $\Delta_P^{NN}$ on the family of convex hulls of the sets in the MAP partition, i.e. $\hat{\mathcal{A}}_n = \big\{\operatorname{conv}\{\mathbf{X}_j \colon j \in J\} \colon J \in \hat{\mathcal{I}}\big\}$, where $\operatorname{conv} A$ is the convex hull of the set $A$. Note that $\hat{\mathcal{A}}_n$ is not necessarily a $P$-partition, since it is possible that $P\Big(\bigcup_{A\in\hat{\mathcal{A}}_n} A\Big) < 1$. Regardless of that fact it is possible to compute $\Delta_P^{NN}(\hat{\mathcal{A}}_n)$, according to the formula (2.60).

**Lemma 3.16.** *Assume that $P$ has bounded support and satisfies $(*)$. Let $X_1, \ldots \sim P$ and let $\hat{\mathcal{A}}_n = \big\{\operatorname{conv}\{\mathbf{X}_j \colon j \in I\} \colon I \in \hat{\mathcal{I}}\big\}$, where $\operatorname{conv} A$ is the convex hull of the set $A$. Then almost surely*

$$\sqrt[n]{Q_{\mathbf{X}_{1:n}}\big(\hat{\mathcal{I}}_n(\mathbf{X}_{1:n})\big)} \simeq \frac{n}{e}\exp\{\Delta_P^{NN}(\hat{\mathcal{A}}_n)\}.$$

*Proof.* From Corollary 3.11 (a) we know $\min\{np_I \colon I \in \hat{\mathcal{I}}_n\} \to \infty$. By applying Stirling formula to each factor $(np_I)!$ and taking into account that by Corollary 3.11 (b) the number of factors is bounded, we obtain that

$$\prod_{I\in\hat{\mathcal{I}}_n}(np_I)! \simeq \prod_{I\in\hat{\mathcal{I}}_n}\Big(\frac{np_I}{e}\Big)^{np_I}\sqrt{2\pi np_I} = \big(\frac{n}{e}\big)^n\sqrt{2\pi n}^{|\hat{\mathcal{I}}_n|-1}\sqrt{\prod_{I\in\hat{\mathcal{I}}_n} p_I}\cdot\left(\prod_{I\in\hat{\mathcal{I}}_n} p_I^{p_I}\right)^n. \tag{3.75}$$

By definition the elements of $\hat{\mathcal{A}}_n$ are convex and hence by Lemma 3.14 the frequencies $p_I$ for $I \in \mathcal{I}_n$ approximate the respective probabilities of sets in $\hat{\mathcal{A}}_n$ uniformly. We also use the fact that the function $\sum p_i \log p_i$ is continuous on the compact set $\triangle^K$ and hence it is uniformly continuous. Hence, as $(|\hat{\mathcal{I}}_n|)_{n=1}^\infty$ is bounded almost surely, it follows that $\sqrt[n]{\prod_{I\in\hat{\mathcal{I}}_n}(np_I)!} \simeq \frac{n}{e}\prod_{I\in\hat{\mathcal{I}}_n} p_I^{p_I} \simeq \frac{n}{e}\prod_{A\in\hat{\mathcal{A}}_n} p_A^{p_A}$. By applying a similar argument to the remaining part of formula (3.1), the result follows by Lemma 3.15 (its assumptions about the probabilities being separated from 0 are satisfied thanks to Corollary 3.11 (a)). □

Now we investigate the convergence of the sequence $\hat{\mathcal{A}}_n$ defined in Lemma 3.16. 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 3.17.** 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 $(X)_\varepsilon = \{x \in \mathbb{R}^d \colon \operatorname{dist}(x, X) < \varepsilon\}$, is a metric on $\mathcal{D}$. It is called the *Hausdorff distance*. The fact that it is a metric follows from Moszyńska, observation 1.2.1.

The two following theorems are crucial for establishing the limits of maximisers. Theorem 3.18 is Theorem 3.2.14 in Moszyńska (2005); it ensures the existence of $\varrho_H$-converging subsequence in every bounded sequence of convex sets. Theorem 3.19 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 the space of closed and convex sets $\mathcal{K}$ is stronger than the symmetric difference metric.

**Theorem 3.18.** *The space $(\mathcal{K}, \varrho_H)$ is finitely compact (i.e. every bounded sequence has a convergent subsequence).*

**Theorem 3.19.** *If $P$ is continuous with respect to the Lebesgue measure then convergence in $\varrho_H$ implies convergence in $d_P$ in the space $\mathcal{K}$.*

Note that the Hausdorff and symmetric difference metrics (cf. Definition 2.28) 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. Lemma A.12 ensures that the desirable properties of compactness are preserved by such extension.

**Definition 3.20.** 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 $\mathcal{A} = \{A^{(1)}, \dots, A^{(k)}\} \in F_K(\mathcal{F})$ and $\mathcal{B} = \{B^{(1)}, \dots, B^{(l)}\} \in F_K(\mathcal{F})$ we define

$$\bar{d}(\mathcal{A}, \mathcal{B}) = \min_{\sigma \in \Sigma_K} \max_{i \leq K} d(A^{(i)}, B^{(\sigma(i))}), \tag{3.76}$$

where $\Sigma_K$ 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.

Now assume that $P$ has bounded support. Then by Theorem 3.18 and Lemma A.12 it follows that $(\hat{\mathcal{A}}_n)_{n=1}^\infty$ has convergent subsequences which have a limit under $\overline{\varrho_H}$ (note that as the support of $P$ is bounded, sets $\hat{\mathcal{A}}$ are also bounded in the $\varrho_H$ metric). Let us denote the (random) set of their limits by $\boldsymbol{E}$. Note that by Theorem 3.18 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 3.16 together with Theorem 3.19 that $\boldsymbol{E}$ consists of finite $P$-partitions that maximise the function $\Delta_P^{NN}$. We state this as Theorem 3.21

**Theorem 3.21.** *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 $\Delta_P^{NN}$.*

*Proof.* Take any $\mathcal{E} = \{E^{(1)}, \ldots, E^{(\tilde{K})}\} \in \mathbf{E}$ and assume that it is a limit of $(\hat{\mathcal{A}}_{n_k})_{k=1}^{\infty}$ in $\overline{\varrho_H}$. By Theorem 3.19 the sequence $(\hat{\mathcal{A}}_{n_k})_{k=1}^{\infty}$ converges to $\mathcal{E}$ also in $\overline{d_P}$. Since for every $k \in \mathbb{N}$ every two sets in the family $\hat{\mathcal{A}}_{n_k}$ are disjoint and hence their intersection has $P$ measure 0. Therefore by the continuity of the intersection with respect to $d_P$ (Doob (1994), Chapter III, Formula (13.3)) we get that $P(E^{(i)} \cap E^{(j)}) = 0$ for $1 \leq i < j \leq \tilde{K}$.

To prove that $\mathcal{E}$ is a $P$-partition it is left to show that $P(\bigcup \mathcal{E}) = 1$ (we denote $\bigcup \mathcal{E} = \bigcup_{E \in \mathcal{E}} E$). Suppose this is not the case. It means that $E_0 = \mathbb{R}^d \setminus \bigcup \mathcal{E}$ is an open set with positive probability. Therefore it includes a ball $B'$ of positive probability. Since $B'$ is a convex set, we get $p_{J_n^{B'}} \to p_{B'} > 0$ and therefore there exist $n' \in \mathbb{N}$ such that $X_{n'} \in B'$. This is not possible, since $X_{n'} \in \bigcup \hat{\mathcal{A}}_n$ for every $n \geq n'$ and therefore $X_{n'} \in \bigcup \mathcal{E}$, which is a contradiction.

By Lemma 3.16 and the continuity of $\Delta_P^{NN}$ with respect to the metric $\overline{d_P}$ we obtain:

$$\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\hat{\mathcal{I}}_n)} \simeq \exp\{\Delta_P^{NN}(\hat{\mathcal{A}}_n)\} \simeq \exp\{\Delta_P^{NN}(\mathcal{E})\}. \tag{3.77}$$

Now take any finite $P$-partition $\mathcal{A}$. We can assume that each $X_n$ belongs to exactly one of the sets in $\mathcal{A}$, $p_{\mathcal{J}_n^A} \to p_A$ and $\overline{\mathbf{X}_{\mathcal{J}_n^A}} \to \mathbb{E}\,(X \,|\, X \in A)$ for $A \in \mathcal{A}$ (it just requires adding a countable number of conditions on the infinite iid sequence with distribution $P$, each of which is satisfied almost surely). By definition of $\hat{\mathcal{I}}_n$ and Lemma 3.13 we get

$$\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\hat{\mathcal{I}}_n)} \geq \sqrt[n]{Q_{\mathbf{X}_{1:n}}(\mathcal{J}_n^{\mathcal{A}})} \simeq \exp\{\Delta_P^{NN}(\mathcal{A})\}. \tag{3.78}$$

Equations (3.77) and (3.78) together give us $\Delta_P^{NN}(\mathcal{E}) \geq \Delta_P^{NN}(\mathcal{A})$ which proves that $\mathcal{E}$ is a finite partition that maximises $\Delta_P^{NN}$.

□

Let $\boldsymbol{M}_\Delta$ be the set of all $P$-partitions that maximise the $\Delta_P^{NN}$ function. Proposition 3.22 states that this set is nonempty and the symmetric distance $d_P$ from $\boldsymbol{M}_\Delta$ to the families of convex hulls $\hat{\mathcal{A}}_n$ of clusters in the MAP clusterings converges to 0.

**Proposition 3.22.** *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.* Let $\mathcal{K}_r$ be the space of all closed and convex subsets of $B(\mathbf{0}, r)$. Note that $\boldsymbol{M}_\Delta$ is closed in $(F_K(\mathcal{K}_r), \overline{d_P})$ as an intersection of the set of maximisers of $\boldsymbol{M}_\Delta$ in $(F_K(\mathcal{K}_r), \overline{d_P})$ and the subspace of $P$-partitions, both of them being closed subspaces of $(F_K(\mathcal{K}_r), \overline{d_P})$. By Theorem 3.21 we know that $\boldsymbol{E} \subseteq \boldsymbol{M}_\Delta$. Now the proof of Proposition 3.22 follows from simple, topological Lemma 3.23, given below. □

**Lemma 3.23.** *Let $(\mathcal{X}, d)$ be a finitely compact metric space, $D \subseteq \mathcal{X}$ a closed set and $(a_n)_{n=1}^{\infty}$ a bounded sequence in $\mathcal{X}$. If every converging subsequence of $(a_n)_{n=1}^{\infty}$ has a limit in $D$ then* $\operatorname{dist}(a_n, D) \to 0$, *where* $\operatorname{dist}(\cdot, \cdot)$ *is the distance function, i.e.*

$$\operatorname{dist}(x, D) = \inf_{y \in D} d(x, y).$$

*Proof.* Suppose that $\limsup \operatorname{dist}(a_n, D) > 0$. Then there exist a subsequence $(a_{n_k})_{k=1}^{\infty}$ and $\varepsilon > 0$ such that $\operatorname{dist}(a_{n_k}, D) > \varepsilon > 0$. This contradicts the fact that $(a_{n_k})_{k=1}^{\infty}$ as a bounded sequence in $\mathcal{X}$ has a converging subsequence whose limit must belong to the closed set $D$. □

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 $\Delta_P^{NN}$ in formula (2.60) 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 3.24 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 $\Sigma$ 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), described in the introduction.

**Theorem 3.24.** *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 $|\hat{\mathcal{I}}_n(\mathbf{X}_{1:n})| > K$.*

*Proof.* Fix $K > 0$. For $\alpha > 0$ let $\Psi(\alpha)$ be defined as

$$\Psi(\alpha) := \inf_{\substack{A \in \mathcal{K}_r \\ P(A) \geq \alpha}} \sup_{\substack{A_1, A_2 \in \mathcal{B} \\ A_1 \cup A_2 = A \\ A_1 \cap A_2 = \emptyset}} P(A_1) \cdot P(A_2) \cdot \|E_P(X \mid X \in A_1) - E_P(X \mid X \in A_2)\|^2 \tag{3.79}$$

In Lemma A.13 we carefully prove an intuitive fact that $\Psi(\alpha) > 0$ for $\alpha > 0$.

We now prove that for $\varepsilon = \frac{1}{8} e \Psi(K^{-1})$ if $\|\Sigma\| < \varepsilon$ then every finite maximiser of the $\Delta_P^{NN}$ function is of size larger than $K$. Take any finite partition $\mathcal{A}$ of $\mathbb{R}^d$ that consists of at most $K$ convex sets with positive $P$ measure. Let $A \in \mathcal{A}$ be the set of the largest probability in $\mathcal{A}$; note that $P(A) \geq K^{-1}$. By definition of $\Psi$ we can divide $A$ into two sets $A_1, A_2$ ($A_1 \cup A_2 = A$, $A_1 \cap A_2 = \emptyset$) such that

$$P(A_1) \cdot P(A_2) \cdot \|E_P(X \mid X \in A_1) - E_P(X \mid X \in A_2)\|^2 > \Psi(K^{-1})/2. \tag{3.80}$$

Let $\mathcal{A}' = \mathcal{A} \cup \{A_1, A_2\} \setminus \{A\}$. Then

$$\begin{aligned}\Delta_P^{NN}(\mathcal{A}') - \Delta_P^{NN}(\mathcal{A}) = \frac{1}{2}\Big(&P(A_1)\|R \cdot E_P(X \mid X \in A_1)\|^2 + P(A_2)\|R \cdot E_P(X \mid X \in A_2)\|^2 - \\ &- P(A)\|R \cdot E_P(X \mid X \in A)\|^2\Big) - P(A_1) \log \frac{1}{P(A_1)} - P(A_2) \log \frac{1}{P(A_2)} + P(A) \log \frac{1}{P(A)}.\end{aligned} \tag{3.81}$$

It is straightforward to verify that $p \log p^{-1} \in [0, \frac{1}{e}]$ for $p \in [0, 1]$ and, since

$$P(A_1) E_P(X \mid X \in A_1) + P(A_2) E_P(X \mid X \in A_2) = P(A) E_P(X \mid X \in A) \tag{3.82}$$

we have

$$P(A_1)\|R\cdot E_P(X\,|\,X\in A_1)\|^2+P(A_2)\|R\cdot E_P(X\,|\,X\in A_2)\|^2-P(A)\|R\cdot E_P(X\,|\,X\in A)\|^2=$$
$$=\frac{P(A_1)P(A_2)}{P(A)}\|R\cdot(E_P(X\,|\,X\in A_1)-E_P(X\,|\,X\in A_2))\|^2. \tag{3.83}$$

Therefore by (3.81) and Lemma A.13 we get

$$\begin{aligned}\Delta_P^{NN}(\mathcal{A}')-\Delta_P^{NN}(\mathcal{A})&\geq\frac{P(A_1)P(A_2)}{P(A)}\|R\cdot(E_P(X\,|\,X\in A_1)-E_P(X\,|\,X\in A_2))\|^2-2e^{-1}\geq\\&\geq\frac{P(A_1)P(A_2)}{P(A)}\frac{1}{\|R^{-1}\|^2}\|E_P(X\,|\,X\in A_1)-E_P(X\,|\,X\in A_2)\|^2-2e^{-1}=\\&=\frac{P(A_1)P(A_2)}{P(A)}\frac{1}{\|\Sigma\|}\|E_P(X\,|\,X\in A_1)-E_P(X\,|\,X\in A_2)\|^2-2e^{-1}\geq\\&\geq\varepsilon^{-1}P(A_1)P(A_2)\|E_P(X\,|\,X\in A_1)-E_P(X\,|\,X\in A_2)\|^2-2e^{-1}\geq\\&\geq\varepsilon^{-1}\Psi(K^{-1})/2-2e^{-1}>2e^{-1}>0.\end{aligned} \tag{3.84}$$

Hence $\mathcal{A}$ is not a maximiser of $\Delta_P^{NN}$ function.

Now let $X_1,X_2,\ldots\overset{\text{iid}}{\sim}P$ and $\hat{\mathcal{A}}_n$ be the family of convex hulls of groups of observations defined by the sequence of the MAP partitions based on $X_1,\ldots,X_n$ (where the MAP partitions were computed in the model with the within group covariance matrix of the norm less than $\varepsilon$). Suppose that there exists a subsequence $(n_i)_{i=1}^{\infty}$ such that $|\hat{\mathcal{A}}_{n_i}|\leq K$ for $i\in\mathbb{N}$. By the compactness of the space $(F_{\tilde{K}}(\mathcal{K}_r),\overline{\varrho_H})$ (cf. Lemma A.12) we get that there is a subsequence $(\hat{\mathcal{A}}_{n_{i_j}})$ that is convergent in this space to a $P$-partition $\mathcal{E}$ of $\mathbb{R}^d$ which is a maximiser of $\Delta_P^{NN}$ (cf. Theorem 3.21). By our previous analysis, $|\mathcal{E}|>K$. On the other hand the probabilities of sets in $\hat{\mathcal{A}}_n$ are separated from 0 (this is a consequence of Corollary 3.10) and this yields a contradiction. □

# Chapter 4

# Normal-Inverse-Wishart with linearly increasing concentration

In Chapter 3, we presented a careful analysis of the Normal-Normal BMM model, where the theoretical covariance structure of each cluster was assumed to be the same and known in advance. The example of a uniform input distribution illustrates that the within-cluster covariance is strongly influenced by the prior covariance parameter. When this 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 (1.37) with $\nu_0 = \alpha + \lambda n$ for parameters $\alpha$ and $\lambda$, while keeping the *expected* within cluster covariance fixed as $\Sigma_0$. More explicitly, we consider the asymptotic limit when, for a sample size $n$, the prior is

$$\Lambda \sim W^{-1}\big(\alpha + \lambda n + d + 1, (\alpha + \lambda n)\Sigma_0\big) \qquad \mu \,|\, \Lambda \sim N\big(\mu_0, \frac{1}{\kappa_0}\Lambda\big). \tag{4.1}$$

This leads to a parametrised family of objectives, which depend on the parameter $\lambda$. For fixed $\Sigma_0$ letting $\lambda$ 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 $\Sigma_0$. When $\lambda > 0$ (inequality strict), we can adapt the methods of the previous chapter (with fixed within-cluster covariance) and prove corresponding results.

The first important result in this chapter is the analogue of the formula (2.73) for the

adjusted Normal-Inverse-Wishart model.

**Proposition 4.1.** *Let $P$ be a probability distribution on $\mathbb{R}^d$ and let $X_1, X_2, \ldots \overset{iid}{\sim} P$. Let $\mathcal{A}$ be a finite $P$-partition. Let $g_{aNIW,n}$ be defined by (1.45) with $\nu_0 = \nu_0(n) = \alpha + \lambda n$, where $\lambda > 0$. Then*

$$\lim_{n\to\infty} \sqrt[n]{Q\big(X_{1:n}, \mathcal{I}_n^{\mathcal{A}}(X_{1:n})\big)} \overset{a.s.}{=} (2\pi)^{-\frac{d}{2}} \exp\{\Delta_{P,\lambda}^{aNIW}(\mathcal{A})\}, \tag{4.2}$$

*where*

$$\begin{aligned}\Delta_{P,\lambda}^{aNIW}(\mathcal{A}) = &-\frac{1}{2}\log|\Sigma_0| - \frac{d}{2} - \\ &- \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big)\log\Big|\frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1}\mathbf{V}_P(X \mid X\in A)\Big| + \mathcal{H}_P(\mathcal{A})\end{aligned} \tag{4.3}$$

The proof follows immediately from Lemma 2.15 and Lemma 4.2.

**Lemma 4.2.** *Let $P$ be a probability distribution on $\mathbb{R}^d$ and let $X_1, X_2, \ldots \overset{iid}{\sim} P$. Let $\mathcal{A}$ be a finite $P$-partition. Let $g_{aNIW,n}$ be given by (1.45), where $\nu_0 = \nu_0(n) = \alpha + \lambda n$ for $\lambda > 0$. Then*

$$\begin{aligned}&\lim_{n\to\infty} \sqrt[n]{g_{aNIW,n}\big(\mathbf{X}_{1:n} \mid \mathcal{I}_n^{\mathcal{A}}(X_{1:n})\big)} \qquad (4.4)\\ &\overset{a.s.}{=} |\Sigma_0|^{-1/2}(2e\pi)^{-d/2} \cdot \prod_{A\in\mathcal{A}} \Big|\frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1}\mathbf{V}_P(X \mid X\in A)\Big|^{-\big(P(A)+\lambda\big)/2}\end{aligned}$$

*Proof of Lemma 4.2.* Let $J_n^A := \{i \le n \colon X_i \in A\}$, for short. For $J \subset [n]$ let $p_J = |J|/n$. Moreover let $\varepsilon_{n,i} = \frac{\alpha+d+1-i}{n}$. By Stirling formula

$$\Gamma_d\left(\frac{|J_n^A| + \nu_0 + d + 1}{2}\right) \overset{\text{a.s.}}{\simeq} \pi^{d(d-1)/4} \prod_{i=0}^{d-1} \left[\sqrt{\pi n(p_{J_n^A} + \lambda + \varepsilon_{n,i})}\left(\frac{n(p_{J_n^A} + \lambda + \varepsilon_{n,i})}{2e}\right)^{\frac{n(p_{J_n^A}+\lambda+\varepsilon_{n,i})}{2}}\right]. \tag{4.5}$$

Hence

$$\sqrt[n]{\Gamma_d\left(\frac{|J_n^A| + \nu_0 + d + 1}{2}\right)} \overset{\text{a.s.}}{\simeq} \prod_{i=0}^{d-1}\left(\frac{n(p_{J_n^A} + \lambda + \varepsilon_{n,i})}{2e}\right)^{\frac{(p_{J_n^A}+\lambda+\varepsilon_{n,i})}{2}}. \tag{4.6}$$

For any fixed $i \le d$

$$\begin{aligned}\prod_{A\in\mathcal{A}}\left(\frac{n(p_{J_n^A} + \lambda + \varepsilon_{n,i})}{2e}\right)^{\frac{(p_{J_n^A}+\lambda+\varepsilon_{n,i})}{2}} &= \\ &= \left(\frac{n}{2e}\right)^{\frac{(1+|\mathcal{A}|\lambda+|\mathcal{A}|\varepsilon_{n,i})}{2}} \prod_{A\in\mathcal{A}}\left(p_{J_n^A} + \lambda + \varepsilon_{n,i}\right)^{\frac{(p_{J_n^A}+\lambda+\varepsilon_{n,i})}{2}} \overset{\text{a.s.}}{\simeq} \qquad (4.7)\\ &\overset{\text{a.s.}}{\simeq} \left(\frac{n}{2e}\right)^{\frac{(1+|\mathcal{A}|\lambda)}{2}} \prod_{A\in\mathcal{A}} (p_A + \lambda)^{\frac{(p_A+\lambda)}{2}},\end{aligned}$$

where we used the fact that $n^{\varepsilon_{n,i}} \to 1$. Thus, by (4.7) and (4.6)

$$\sqrt[n]{\prod_{A\in\mathcal{A}} \Gamma_d\left(\frac{|J_n^A| + \nu_0 + d + 1}{2}\right)} \overset{\text{a.s.}}{\simeq} \left(\frac{n}{2e}\right)^{\frac{d(1+|\mathcal{A}|\lambda)}{2}} \prod_{A\in\mathcal{A}} (p_A + \lambda)^{\frac{d(p_A+\lambda)}{2}}, \tag{4.8}$$

Similarly, using $n^{\varepsilon_{n,i}} \to 1$ again we can show that

$$\sqrt[n]{\Gamma_d\left(\frac{\nu_0 + d + 1}{2}\right)} \simeq \prod_{i=0}^{d-1} \left(\frac{n(\lambda + \varepsilon_{n,i})}{2e}\right)^{\frac{(\lambda+\varepsilon_{n,i})}{2}} \simeq \left(\frac{n\lambda}{2e}\right)^{\frac{(\lambda d)}{2}} \tag{4.9}$$

Also, the following statement hold

$$\sqrt[n]{|\nu_0\Sigma_0|^{(\nu_0+d+1)/2}} \simeq (n\lambda)^{d\lambda/2}|\Sigma_0|^{\lambda/2} \tag{4.10}$$

From (4.9) and (4.10) we get

$$\sqrt[n]{\frac{|\nu_0\Sigma_0|^{(\nu_0+d+1)/2}}{\Gamma_d\left(\frac{\nu_0+d+1}{2}\right)}} \simeq \left(\frac{|\Sigma_0|}{(2e)^d}\right)^{\lambda/2}. \tag{4.11}$$

Clearly

$$\lim_{n\to\infty} \sqrt[n]{\frac{\kappa_0}{\kappa_0 + |J_n^A|}} = 1. \tag{4.12}$$

By the Strong Law of Large Numbers we have that

$$\lim_{n\to\infty} \frac{1}{|J_n^A|} \sum_{i\in J_n^A} (x_i - \overline{\boldsymbol{x}_A})(x_i - \overline{\boldsymbol{x}_A})^\top = \mathbf{V}_P(X \,|\, X \in A) \quad \text{a.s.} \quad \text{for } A \in \mathcal{A} \tag{4.13}$$

and hence, recalling the definition (1.47) of $\Sigma(\boldsymbol{x})$, for $A \in \mathcal{A}$

$$\begin{aligned}\left|\Sigma(\mathbf{X}_{J_n^A})\right| &= |J_n^A|^d \cdot \left|\frac{\nu_0\Sigma_0}{|J_n^A|} + \frac{\sum_{i\in J_n^A}(x_i - \overline{\boldsymbol{x}_A})(x_i - \overline{\boldsymbol{x}_A})^\top}{|J_n^A|} + \frac{k_0(\overline{\boldsymbol{x}_A} - \mu_0)(\overline{\boldsymbol{x}_A} - \mu_0)^\top}{k_0 + |J_n^A|}\right| = \\ &= n^d p_{J_n^A}^d \cdot \left|\frac{\nu_0\Sigma_0}{|J_n^A|} + \frac{\sum_{i\in J_n^A}(x_i - \overline{\boldsymbol{x}_A})(x_i - \overline{\boldsymbol{x}_A})^\top}{|J_n^A|} + \frac{k_0(\overline{\boldsymbol{x}_A} - \mu_0)(\overline{\boldsymbol{x}_A} - \mu_0)^\top}{k_0 + |J_n^A|}\right|\end{aligned} \tag{4.14}$$

Let

$$\hat{V}_{A,n} = \frac{\nu_0\Sigma_0}{|J_n^A|} + \frac{\sum_{i\in J_n^A}(x_i - \overline{\boldsymbol{x}_A})(x_i - \overline{\boldsymbol{x}_A})^\top}{|J_n^A|} + \frac{k_0(\overline{\boldsymbol{x}_A} - \mu_0)(\overline{\boldsymbol{x}_A} - \mu_0)^\top}{k_0 + |J_n^A|} \tag{4.15}$$

Then for $\varepsilon_n = \frac{d+1+\alpha}{n}$,

$$\sqrt[n]{\left|\Sigma(\mathbf{X}_{J_n^A})\right|^{(|J_n^A|+\nu_0+d+1)/2}} = (n^d p_{J_n^A}^d |\hat{V}_{A,n}|)^{(p_{J_n^A}+\lambda+\varepsilon_n)/2}. \tag{4.16}$$

It follows that

$$\sqrt[n]{\prod_{A\in\mathcal{A}} \left|\Sigma(\mathbf{X}_{J_n^A})\right|^{(|J_n^A|+\nu_0+d+1)/2}} = n^{d(1+|\mathcal{A}|\lambda+|\mathcal{A}|\varepsilon_n)} \prod_{A\in\mathcal{A}} (p_{J_n^A}^d |\hat{V}_{A,n}|)^{(p_{J_n^A}+\lambda+\varepsilon_n)/2}. \tag{4.17}$$

Note that by the Strong Law of Large Numbers

$$\lim_{n\to\infty} (p^d_{J^A_n}|\hat{V}_{A,n}|)^{(p_{J^A_n}+\lambda+\varepsilon_n)/2} = P(A)^d \left(\left|\frac{\lambda}{P(A)}\Sigma_0 + \mathbf{V}_P(X\,|\,X\in A)\right|\right)^{(P(A)+\lambda)/2}. \quad (4.18)$$

Using (4.17) and (4.18), together with the fact that $n^{\varepsilon_n} \to 1$, we obtain

$$\begin{aligned}\prod_{A\in\mathcal{A}} \sqrt[n]{|\Sigma(\mathbf{X}_{J^A_n})|^{(|J^A_n|+\nu_0+d+1)/2}} &\overset{\text{a.s.}}{\simeq} n^{d(1+|\mathcal{A}|\lambda)} \prod_{A\in\mathcal{A}} \left(P(A)^d \left|\frac{\lambda}{P(A)}\Sigma_0 + \mathbf{V}_P(X\,|\,X\in A)\right|\right)^{(P(A)+\lambda)/2} = \\ &= n^{d(1+|\mathcal{A}|\lambda)} \prod_{A\in\mathcal{A}} \left((P(A)+\lambda)^d \left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathbf{V}_P(X\,|\,X\in A)\right|\right)^{(P(A)+\lambda)/2} = \\ &= n^{d(1+|\mathcal{A}|\lambda)}|\Sigma_0|^{(1+|\mathcal{A}|\lambda)/2} \prod_{A\in\mathcal{A}} \left((P(A)+\lambda)^d \left|\frac{\lambda}{P(A)+\lambda}I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1}\mathbf{V}_P(X\,|\,X\in A)\right|\right)^{(P(A)+\lambda)/}\end{aligned} \quad (4.19)$$

The proof follows from plugging (4.8), (4.11), (4.12) and (4.19) into (1.45). □

Let us highlight important differences between the formulas (4.3) (the objective to be maximised when $\lambda > 0$) and (2.73) (the limiting objective as $\lambda \to 0$). Equation (4.3) may be expressed equivalently as:

$$\begin{aligned}\Delta^{aNIW}_{P,\lambda}(\mathcal{A}) = \frac{1}{2}|\mathcal{A}|\cdot\lambda\log|\Sigma_0| - \frac{d}{2} - \\ - \frac{1}{2}\sum_{A\in\mathcal{A}}\big(P(A)+\lambda\big)\log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathbf{V}_P(X\,|\,X\in A)\right| + \sum_{A\in\mathcal{A}} P(A)\log P(A).\end{aligned} \quad (4.20)$$

When $\lambda > 0$, the 'variance' part now contains a convex combination of the true within-cluster covariance matrix $\mathbf{V}_P(X\,|\,X\in A)$ and the apriori expected value of the within-cluster covariance $\Sigma_0$.

For $\lambda = 0$, the log-determinant is multiplied by the probability $P(A)$, while for $\lambda > 0$, this multiplier is $(P(A)+\lambda)$ and the sum over all $A \in \mathcal{A}$ is not an expected value, unlike for $\lambda = 0$,. The general formula also has a term that depends linearly on the number of clusters in the partition $\mathcal{A}$, which disappears when $\lambda = 0$.

The quantity $\frac{1}{2}\log|\Sigma_0|$, is independent of the partition $\mathcal{A}$ and is therefore irrelevant for the problem of finding a maximiser.

We now consider what happens to the objective (4.3) when $\lambda \to \infty$.

**Proposition 4.3.** *Let $\mathcal{A}$ be any partition of $\mathbb{R}^d$ and $P$ a probability measure. Then, up to constants (i.e. differences which do not depend on the partition)*

*(a) for fixed $\mathcal{A}$ the function $\lambda \mapsto \Delta_{P,\lambda}(\mathcal{A})$ is decreasing;*

*(b)* $\lim_{\lambda\to\infty}\Delta_{P,\lambda}(\mathcal{A}) = -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}\sum_{A\in\mathcal{A}} P(A)\mathrm{tr}(\Sigma_0^{-1}\mathbf{V}_P(X\,|\,X\in A)) + \sum_{A\in\mathcal{A}} P(A)\log P(A)$.

*Proof of Proposition 4.3.* We start with a crucial lemma.

**Lemma 4.4.** *Let $x_0 > 0$ and let $\Sigma$ be a symmetric, $d \times d$ matrix such that $I_d + \frac{\Sigma}{x}$ is positive definite for all $x \geq x_0$. Then*

*(i)* $x \mapsto x \log \left| I_d + \frac{\Sigma}{x} \right|$ *is an increasing function on* $(x_0, \infty)$

*(ii)* $\lim_{x\to\infty} x \log \left| I_d + \frac{\Sigma}{x} \right| = \mathrm{tr}(\Sigma)$

*Proof.* Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of the matrix $\Sigma$. Then the eigenvalues of the matrix $I_d + \frac{\Sigma}{x}$ are $(1 + \frac{\lambda_i}{x}) > 0$ for $i \leq d$ and $x \geq x_0$. Therefore

$$\lim_{x\to\infty} x \log \left| I_d + \frac{\Sigma}{x} \right| = \lim_{x\to\infty} \sum_{i=1}^{d} x \log \left( 1 + \frac{\lambda_i}{x} \right) = \sum_{i=1}^{d} \lambda_i = \mathrm{tr}(\Sigma). \tag{4.21}$$

The fact that $x \mapsto x \log \left| I_d + \frac{\Sigma}{x} \right|$ is increasing follows from the fact that each of the functions $x \log \left( 1 + \frac{\lambda_i}{x} \right)$ is increasing (regardless of the sign of $\lambda_i$). □

This is key to establishing Proposition 4.3. Firstly note that

$$\begin{aligned}
&\left| \frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda} \Sigma_0^{-1} \mathbf{V}_P(X \mid X \in A) \right| = \\
&\qquad = |\Sigma_0^{-1}| \left| \frac{\lambda}{P(A)+\lambda} \Sigma_0 + \frac{P(A)}{P(A)+\lambda} \mathbf{V}_P(X \mid X \in A) \right| = \\
&\qquad = |\Sigma_0^{-1/2}| \left| \frac{\lambda}{P(A)+\lambda} \Sigma_0 + \frac{P(A)}{P(A)+\lambda} \mathbf{V}_P(X \mid X \in A) \right| |\Sigma_0^{-1/2}| = \\
&\qquad = \left| \frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda} \Sigma_0^{-1/2} \mathbf{V}_P(X \mid X \in A) \Sigma_0^{-1/2} \right|
\end{aligned} \tag{4.22}$$

By Lemma 4.4 we have that

$$\begin{aligned}
&\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log \left| \frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda} \Sigma_0^{-1} \mathbf{V}_P(X \mid X \in A) \right| = \\
&\qquad = \sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log \left| I_d + \frac{P(A)}{P(A)+\lambda} (\Sigma_0^{-1/2} \mathbf{V}_P(X \mid X \in A) \Sigma_0^{-1/2} - I_d) \right|
\end{aligned} \tag{4.23}$$

is increasing in $\lambda$ (plug $x = P(A) + \lambda$, $x_0 = P(A)$, $\Sigma = \Sigma_0^{-1/2} \mathbf{V}_P(X \mid X \in A) \Sigma_0^{-1/2} - I_d$) and

$$\begin{aligned}
&\lim_{\lambda\to\infty} \sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log \left| \frac{\lambda}{P(A)+\lambda} I_d + \frac{P(A)}{P(A)+\lambda} \Sigma_0^{-1} \mathbf{V}_P(X \mid X \in A) \right| = \\
&\quad = \lim_{\lambda\to\infty} \sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log \left| I_d + \frac{P(A)}{P(A)+\lambda} (\Sigma_0^{-1/2} \mathbf{V}_P(X \mid X \in A) \Sigma_0^{-1/2} - I_d) \right| = \\
&\quad = \sum_{A\in\mathcal{A}} \mathrm{tr}\Big( P(A) (\Sigma_0^{-1/2} \mathbf{V}_P(X \mid X \in A) \Sigma_0^{-1/2} - I_d) \Big) = \\
&\quad = \sum_{A\in\mathcal{A}} P(A) \mathrm{tr}(\Sigma_0^{-1} \mathbf{V}_P(X \mid X \in A)) - d,
\end{aligned} \tag{4.24}$$

which concludes the proof of Proposition 4.3. □

Note that part (b) of Proposition 4.3 gives (up to a constant) the same objective to be maximised as $\Delta_P^{NN}$ (cf. (2.63)). Therefore, as $\lambda$ ranges from $\infty$ down to 0, we have a continuous transition between the fixed within-cluster covariance $\Sigma_0$ and a classifier which does not include any prior information at all on the within-cluster covariance; the $\lambda$ parameter gives a continuous transition between the fixed covariance Normal-Normal and the Normal-Inverse-Wishart models.

**Proposition 4.5.** *Let $\mathcal{H}_P(\mathcal{A})$ be defined as in (2.48). If $\mathbf{V}_P(X) < \infty$ then for any $\lambda > 0$ the function $\Delta_{P,\lambda}^{aNIW}(\mathcal{A}) - \mathcal{H}(\mathcal{A})$ is a bounded function on the space of all $P$-partitions of $\mathbb{R}^d$.*

*Proof of Proposition 4.5.* For any $P$-partition $\mathcal{A}$ of the observation space (cf. Lemma A.14) we have

$$\sum_{A\in\mathcal{A}} P(A)\mathbf{V}_P(X \mid X \in A) \preceq \mathbf{V}_P(X), \tag{4.25}$$

where $\preceq$ is the Löwner partial order, i.e. $A \preceq B$ if and only if $B - A$ is nonnegative definite. If $A \preceq B$ then $CAC^\top \preceq CBC^\top$ for any positive definite matrix $C$. Since $\Sigma_0^{-1/2}$ is positive definite and symmetric, we get

$$\sum_{A\in\mathcal{A}} P(A)\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2} \preceq \Sigma_0^{-1/2}\mathbf{V}_P(X)\Sigma_0^{-1/2}, \tag{4.26}$$

Since the trace of a matrix is increasing with respect to the Löwner partial order, (4.26) yields

$$\sum_{A\in\mathcal{A}} P(A)\mathrm{tr}(\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2}) \leq \mathrm{tr}(\Sigma_0^{-1/2}\mathbf{V}_P(X)\Sigma_0^{-1/2}). \tag{4.27}$$

Let $\omega_{A,i}$ be the $i$-th largest eigenvalue of the matrix $\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2}$. Clearly $\omega_{A,i} \geq 0$. Then by (4.27) we get

$$\sum_{A\in\mathcal{A}}\sum_{i=1}^{d} P(A)\omega_{A,i} \leq \mathrm{tr}(\Sigma_0^{-1/2}\mathbf{V}_P(X)\Sigma_0^{-1/2}). \tag{4.28}$$

Note that the eigenvalues of $\frac{\lambda}{P(A)+\lambda}I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2}$ are $\frac{\lambda}{P(A)+\lambda} + \frac{P(A)}{P(A)+\lambda}\omega_{A,i}$ and hence for every $A \in \mathcal{A}$

$$\det\left(\frac{\lambda}{P(A)+\lambda}I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2}\right) = \prod_{i=1}^{d}\left(\frac{\lambda}{P(A)+\lambda} + \frac{P(A)}{P(A)+\lambda}\omega_{A,i}\right). \tag{4.29}$$

By (4.29)

$$\log\left|\frac{\lambda}{P(A)+\lambda}I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1/2}\mathbf{V}_P(X \mid X \in A)\Sigma_0^{-1/2}\right| = \sum_{i=1}^{d}\log\left(\frac{\lambda}{P(A)+\lambda} + \frac{P(A)}{P(A)+\lambda}\omega_{A,i}\right). \tag{4.30}$$

For every $x > -1$ we have $\log(1+x) \geq x - \frac{x^2}{2}$, hence

$$\log\left(\frac{\lambda}{P(A)+\lambda} + \frac{P(A)}{P(A)+\lambda}\omega_{A,i}\right) \geq \frac{P(A)}{P(A)+\lambda}(\omega_{A,i}-1) - \frac{1}{2}\left(\frac{P(A)}{P(A)+\lambda}\right)^2(\omega_{A,i}-1)^2. \tag{4.31}$$

Since $\omega_{A,i} \geq 0$ and $\sum_{A\in\mathcal{A}} P(A) = 1$, we get

$$\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}(P(A)+\lambda)\frac{P(A)}{P(A)+\lambda}(\omega_{A,i}-1) = \sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)(\omega_{A,i}-1) \geq -d. \tag{4.32}$$

On the other hand

$$\begin{aligned}
\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}(P(A)+\lambda)\left(\frac{P(A)}{P(A)+\lambda}\right)^2(\omega_{A,i}-1)^2 &= \sum_{A\in\mathcal{A}}\sum_{i=1}^{d}\frac{P(A)^2}{P(A)+\lambda}(\omega_{A,i}-1)^2 \leq \\
&\leq \frac{1}{\lambda}\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)^2(\omega_{A,i}-1)^2 = \frac{1}{\lambda}\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)^2(\omega_{A,i}^2-2\omega_{A,i}+1) \leq \\
&\leq \frac{1}{\lambda}\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)^2(\omega_{A,i}^2+1) \leq \frac{1}{\lambda}\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)^2\omega_{A,i}^2 + \frac{d}{\lambda} \leq \\
&\leq \frac{1}{\lambda}\left(\sum_{A\in\mathcal{A}}\sum_{i=1}^{d}P(A)\omega_{A,i}\right)^2 + \frac{d}{\lambda} \overset{(4.28)}{\leq} \frac{1}{\lambda}\mathrm{tr}(\Sigma_0^{-1/2}\mathbf{V}_P(X)\Sigma_0^{-1/2})^2 + \frac{d}{\lambda}.
\end{aligned} \tag{4.33}$$

By joining (4.30) with (4.31), (4.32) and (4.33), we get that

$$\begin{aligned}
\sum_{A\in\mathcal{A}}(P(A)+\lambda)\log\left|\frac{\lambda}{P(A)+\lambda}I_d + \frac{P(A)}{P(A)+\lambda}\Sigma_0^{-1/2}\mathbf{V}_P(X\,|\,X\in A)\Sigma_0^{-1/2}\right| \geq \\
\geq -d - \frac{1}{2\lambda}\left(\mathrm{tr}(\Sigma_0^{-1/2}\mathbf{V}_P(X)\Sigma_0^{-1/2})^2 + d\right).
\end{aligned} \tag{4.34}$$

Which, together with an obvious inequality $\sum_{A\in\mathcal{A}} P(A)\log P(A) \leq 0$ and (4.22), finishes the proof. □

## 4.1 Linear growth of clusters

In Chapter 3 we considered Normal-Normal model, where the within-cluster covariance $\Sigma_0$ is fixed and known. We showed that if the prior on the space of partition is the Chinese Restaurant Process (cf. (1.14)) and the data sequence is bounded then the size of the smallest cluster grows linearly with the number of observations and hence the number of clusters in the MAP partition is bounded from above. In this section we show that the techniques from Chapter 3 can be used to prove the linear growth of cluster size in the model (4.1) when $\lambda > 0$. This is encapsulated in the following result:

**Proposition 4.6.** *Let $\boldsymbol{x} = (x_n)_{n=1}^{\infty}$ be a bounded infinite sequence of points in $\mathbb{R}^d$. Consider the model (4.1) and the standard Chinese Restaurant prior on the space of partitions. Let $\hat{\mathcal{I}}_n$ be the MAP partition of $\mathbf{x}_{1:n}$. Then*

$$\liminf_{n\to\infty}\min_{I\in\hat{\mathcal{I}}_n}\frac{|I|}{n} > 0. \tag{4.35}$$

**Corollary 4.7.** *With the assumptions and notation of Proposition 4.6 the number of clusters in the MAP partition is almost surely bounded, i.e.*

$$\limsup_{n\to\infty} |\hat{\mathcal{I}}_n| < \infty. \tag{4.36}$$

We now prove Proposition 4.6.

### 4.1.1 Proof of Proposition 4.6

We present some preliminaries; the proof is given later in this subsection.

#### Dealing with the maximal cluster

The first step to prove Proposition 4.6 is showing that in the MAP partition the size of the *maximal* cluster is proportional to the number of observations.

**Lemma 4.8.** *With the assumptions and notation of Proposition 4.6 we have*

$$\liminf_{n\to\infty} \max_{I\in\hat{\mathcal{I}}_n} |I|/n > 0. \tag{4.37}$$

*Proof.* The proof requires Lemma 4.11, which is stated and proved below. This presents a comparison with the single cluster partition (all observations belong to the same cluster) and shows that for any sequence of partitions where $\min_{I\in\hat{I}_n} |I|/n$ goes to 0, the single cluster partition gives (asymptotically) a larger value for the objective.

Lemma 4.8 is then a straightforward consequence of this, since if (4.37) is not satisfied then the posterior probability of $\hat{\mathcal{I}}_n$ is less than the posterior probability of $\{[n]\}$ (all $n$ observations belong to a single cluster) for sufficiently large $n$, contradicting the definition of the MAP partition. □

Lemma 4.11 is quite subtantial and to prove it, we need some additional lemmas.

We first establish an upper bound.

**Lemma 4.9.** *Let $(x_n)_{n=1}^{\infty}$ be an infinite sequence of points in $\mathbb{R}^d$. Let $g_{aNIW,n}$ be defined by (1.45), with $\nu_0 = \alpha + \lambda n$, where $\lambda > 0$ (as in Proposition 4.6). Let $(\mathcal{I}_n)_{n=1}^{\infty}$ be any sequence of partitions. Then*

$$\limsup_{n\to\infty} \sqrt[n]{g_{aNIW,n}(\mathbf{x}_{1:n} \,|\, \mathcal{I}_n)} < \infty. \tag{4.38}$$

*Proof.* Let us rewrite (1.45), bearing in mind that $\nu_0 = \alpha + \lambda n$. To improve notation, let

$\beta = \alpha + d + 1$.

$$g_{aNIW,n}(\boldsymbol{x}_{1:n} \,|\, \mathcal{I}_n) = \pi^{-\frac{dn}{2}} \left( \frac{|(\alpha+\lambda n)\Sigma_0|^{\frac{\beta+\lambda n}{2}} \kappa_0^{\frac{d}{2}}}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)} \right)^{|\mathcal{I}_n|} \cdot \\ \cdot \prod_{I \in \mathcal{I}_n} \Gamma_d\left(\frac{\beta+\lambda n+|I|}{2}\right) (\kappa_0+|I|)^{-\frac{d}{2}} \det\left(\Sigma(\boldsymbol{x}_I)\right)^{-\frac{\beta+\lambda n+|I|}{2}}, \tag{4.39}$$

where

$$\Sigma(\boldsymbol{u}) = (\alpha+\lambda n)\Sigma_0 + \sum_{i=1}^{k} (u_i - \overline{\boldsymbol{u}})(u_i - \overline{\boldsymbol{u}})^\top + \frac{\kappa_0 k}{\kappa_0 + k} (\overline{\boldsymbol{u}} - \mu_0)(\overline{\boldsymbol{u}} - \mu_0)^\top. \tag{4.40}$$

Then (4.39) can be written as:

$$g_{aNIW,n}(\boldsymbol{x}_{1:n} \,|\, \mathcal{I}_n) = \pi^{-\frac{dn}{2}} \prod_{I \in \mathcal{I}} \frac{\Gamma_d\left(\frac{\beta+\lambda n+|I|}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)} \left(\frac{\kappa_0}{\kappa_0+|I|}\right)^{\frac{d}{2}} \left|\frac{\Sigma_0^{-1}}{\alpha+\lambda n}\Sigma(\boldsymbol{x}_I)\right|^{-\frac{\beta+\lambda n}{2}} \left|\Sigma(\boldsymbol{x}_I)\right|^{-\frac{|I|}{2}}. \tag{4.41}$$

From Equation (A.65) we have for every $I \in \mathcal{I}_n$

$$\frac{\Gamma_d\left(\frac{\beta+\lambda n+|I|}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)} < \left(\frac{\beta+\lambda n+|I|}{2}\right)^{\frac{|I|d}{2}}. \tag{4.42}$$

Note that

$$\Sigma(\boldsymbol{u}) \succeq (\alpha+\lambda n)\Sigma_0. \tag{4.43}$$

By (4.42) and (4.43)

$$g_{aNIW,n}(\boldsymbol{x}_{1:n} \,|\, \mathcal{I}_n) \leq \pi^{-\frac{dn}{2}} \prod_{I \in \mathcal{I}} \left(\frac{\beta+\lambda n+|I|}{2}\right)^{\frac{|I|d}{2}} \left|(\alpha+\lambda n)\Sigma_0\right|^{-\frac{|I|}{2}} = \\ = \left(\frac{\pi}{2}\right)^{-\frac{dn}{2}} \prod_{I \in \mathcal{I}} \left|\frac{\alpha+\lambda n}{\beta+\lambda n+|I|}\Sigma_0\right|^{-\frac{|I|}{2}} < \left(\frac{\pi}{2}\right)^{-\frac{dn}{2}} \left|\frac{\lambda}{\beta+\lambda+1}\Sigma_0\right|^{-\frac{n}{2}}. \tag{4.44}$$

Hence

$$\sqrt[n]{g_{aNIW,n}(\boldsymbol{x}_{1:n} \,|\, \mathcal{I}_n)} < \left(\frac{\pi}{2}\right)^{-\frac{d}{2}} \left|\frac{\lambda}{\beta+\lambda+1}\Sigma_0\right|^{-\frac{1}{2}}, \tag{4.45}$$

which is finite, thus establishing the lemma. □

We now establish a lower bound for the single cluster partition.

**Lemma 4.10.** *With the assumptions of Proposition 4.6*

$$\liminf_{n\to\infty} \sqrt[n]{g_{aNIW,n}(\mathbf{x}_{1:n} \,|\, \{[n]\})} > 0. \tag{4.46}$$

*Proof.* Equation (4.41) with the single cluster $\mathcal{I}_n = \{[n]\}$ gives:

$$g_{aNIW,n}(\boldsymbol{x}_{1:n} \,|\, \{[n]\}) = \pi^{-\frac{dn}{2}} \frac{\Gamma_d\left(\frac{\beta+\lambda n+n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)} \left(\frac{\kappa_0}{\kappa_0+n}\right)^{\frac{d}{2}} \left|\frac{\Sigma_0^{-1}}{\alpha+\lambda n}\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{\beta+\lambda n}{2}} \left|\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{n}{2}}. \tag{4.47}$$

Note that

$$\lim_{n\to\infty} \sqrt[n]{\left(\frac{\kappa_0}{\kappa_0+n}\right)^{\frac{d}{2}}} = \lim_{n\to\infty} \exp\left\{-\frac{d}{2n}\log\left(1+\frac{n}{\kappa_0}\right)\right\} = 1. \tag{4.48}$$

Assume that $\|x_i\| \leq r$ for some $r > 0$ and all $i \in \mathbb{N}$. By Lemma A.8, Lemma A.9 and the triangle inequality

$$|\Sigma(\boldsymbol{x}_{1:n})/n| \leq \|\Sigma(\boldsymbol{x}_{1:n})/n\|^d \leq \left(\left\|\frac{\alpha+\lambda n}{n}\Sigma_0\right\| + (2r)^2 + \frac{\kappa_0}{n}(2r)^2\right)^d \tag{4.49}$$

and hence

$$\limsup_{n\to\infty} |\Sigma(\boldsymbol{x}_{1:n})/n| \leq \left(\left\|\lambda\Sigma_0\right\| + (2r)^2\right)^d. \tag{4.50}$$

As a consequence of (4.50)

$$\begin{aligned}\liminf_{n\to\infty} \sqrt[n]{\left|\frac{\Sigma_0^{-1}}{\alpha+\lambda n}\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{\beta+\lambda n}{2}}} &\geq \left(\limsup_{n\to\infty}\left|\frac{n\Sigma_0^{-1}}{\alpha+\lambda n}\right|\right)^{-\frac{\lambda}{2}} \left(\limsup_{n\to\infty}\left|\Sigma(\boldsymbol{x}_{1:n})/n\right|\right)^{-\frac{\lambda}{2}} = \\ &= \left|\frac{\Sigma_0^{-1}}{\lambda}\right|^{-\frac{\lambda}{2}} \left(\limsup_{n\to\infty}\left|\Sigma(\boldsymbol{x}_{1:n})/n\right|\right)^{-\frac{\lambda}{2}} > 0.\end{aligned} \tag{4.51}$$

Using Equation (A.65)

$$\begin{aligned}\frac{\Gamma_d\left(\frac{\beta+\lambda n+n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)}\left|\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{n}{2}} &> \left(\frac{\beta+\lambda n+n-d}{2e}\right)^{\frac{nd}{2}} \left|\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{n}{2}} = \\ &= \left(\frac{\beta+\lambda n+n-d}{2ne}\right)^{\frac{nd}{2}} \left|\Sigma(\boldsymbol{x}_{1:n})/n\right|^{-\frac{n}{2}}\end{aligned} \tag{4.52}$$

and hence, again using (4.50)

$$\liminf_{n\to\infty} \sqrt[n]{\frac{\Gamma_d\left(\frac{\beta+\lambda n+n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)}\left|\Sigma(\boldsymbol{x}_{1:n})\right|^{-\frac{n}{2}}} \geq \left(\frac{\lambda+1}{2e}\right)^{\frac{d}{2}} \left(\limsup_{n\to\infty}\left|\Sigma(\boldsymbol{x}_{1:n})/n\right|^{-\frac{1}{2}}\right) > 0. \tag{4.53}$$

Plugging (4.48), (4.51) and (4.53) into (4.47) establishes the result. □

**Lemma 4.11.** *Under the assumptions of Proposition 4.6, if $(\mathcal{I}_n)_{n=1}^{\infty}$ is a sequence of partitions such that*

$$\liminf_{n\to\infty} \max_{I\in\mathcal{I}_n} |I|/n = 0 \tag{4.54}$$

*then*

$$\liminf_{n\to\infty} \sqrt[n]{\frac{Q(\mathcal{I}_n, \boldsymbol{u}_{1:n})}{Q(\{[n]\}, \boldsymbol{u}_{1:n})}} = 0. \tag{4.55}$$

*($Q(\cdot,\cdot)$ is the joint probability of observations and their partition, given by (1.8)).*

*Proof.* From the definition,

$$\frac{Q(\mathcal{I}_n, \boldsymbol{u}_{1:n})}{Q(\{[n]\}, \boldsymbol{u}_{1:n})} = \frac{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \mathcal{I}_n)\mathcal{P}_{\pi,n}(\mathcal{I}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, [n])\mathcal{P}_{\pi,n}(\{[n]\})}. \tag{4.56}$$

From Lemma 4.9 and Lemma 4.10, the denominator of (4.57) is bounded from below and the numerator bounded from above; hence:

$$\limsup_{n\to\infty} \sqrt[n]{\frac{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \mathcal{I}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, [n])}} < \infty. \tag{4.57}$$

For the Chinese Restaurant Process:

$$\frac{\mathcal{P}_{\pi,n}(\mathcal{I}_n)}{\mathcal{P}_{\pi,n}(\{[n]\})} = \frac{\alpha^{|\mathcal{I}_n|} n}{\alpha \prod_{I\in\mathcal{I}_n} |I|} \frac{\prod_{I\in\mathcal{I}_n} |I|!}{n!} \tag{4.58}$$

By Lemma 3.6, under the assumption (4.54) we get

$$\liminf_{n\to\infty} \sqrt[n]{\frac{\prod_{I\in\mathcal{I}_n} |I|!}{n!}} = 0. \tag{4.59}$$

Moreover

$$\limsup_{n\to\infty} \sqrt[n]{\frac{\alpha^{|\mathcal{I}_n|} n}{\alpha \prod_{I\in\mathcal{I}_n} |I|}} \leq \limsup_{n\to\infty} \sqrt[n]{\frac{\alpha^{|\mathcal{I}_n|} n}{\alpha}} \leq \alpha \tag{4.60}$$

(by considering the 'worst case' $|\mathcal{I}_n| = n$). By plugging (4.59) and (4.60) into (4.58), we get

$$\liminf_{n\to\infty} \sqrt[n]{\frac{\mathcal{P}_{\pi,n}(\mathcal{I}_n)}{\mathcal{P}_{\pi,n}(\{[n]\})}} = 0. \tag{4.61}$$

The inequality (4.61) together with (4.57), applied to (4.56) concludes the proof. □

### Proof of Proposition 4.6

We are now in a position to give the proof. Let $\beta = \alpha + d + 1$. Suppose that the smallest cluster of the maximising partition grows sub-linearly in $n$. Let $\tilde{\mathcal{I}}_n$ be the partition obtained by joining the smallest and the largest cluster; let us denote the sizes of these clusters by $m_n$ and $M_n$. By the assumptions $\liminf_{n\to\infty} m_n/n = 0$ and by Lemma 4.8 $\liminf_{n\to\infty} M_n/n =: \gamma > 0$. Assume that the sequence $\boldsymbol{x}$ is contained in a ball of radius $r$, centered at the origin, i.e. $\|x_i\| \leq r$ for all $i \in \mathbb{N}$. The ratio $\frac{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \hat{\mathcal{I}}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \tilde{\mathcal{I}}_n)}$ may be written as:

$$\frac{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \hat{\mathcal{I}}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n} \,|\, \tilde{\mathcal{I}}_n)} = \mathtt{A}_n \mathtt{B}_n \mathtt{C}_n \mathtt{D}_n \mathtt{E}_n, \tag{4.62}$$

where

$$\mathtt{A}_n = \left(\frac{\kappa_0(\kappa_0+m_n+M_n)}{(\kappa_0+m_n)(\kappa_0+M_n)}\right)^{\frac{d}{2}}, \ \mathtt{B}_n = \frac{\Gamma_d\left(\frac{\beta+\lambda n+m_n}{2}\right)\Gamma_d\left(\frac{\beta+\lambda n+M_n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)\Gamma_d\left(\frac{\beta+\lambda n+m_n+M_n}{2}\right)},$$

$$\mathtt{C}_n = \left(\frac{|(\alpha+\lambda n)\Sigma_0|}{|\Sigma(\boldsymbol{x}_{I_{\min}})|}\right)^{(\beta+\lambda n)/2}, \ \mathtt{D}_n = \left(\frac{|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})|}{|\Sigma(\boldsymbol{x}_{I_{\min}})|}\right)^{m_n/2}, \ \mathtt{E}_n = \left(\frac{|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})|}{|\Sigma(\boldsymbol{x}_{I_{\max}})|}\right)^{(\beta+\lambda n+M_n)/2} \tag{4.63}$$

Clearly $\lim_{n\to\infty} m_n/M_n = 0$ and, from (1.47) with $\nu_0 = \alpha+\lambda n$, $(\alpha+\lambda n)\Sigma_0 \preceq \Sigma(\boldsymbol{x}_{I_{\min}})$, so

$$\limsup_{n\to\infty} \mathtt{A}_n \le 1 \quad \text{and} \quad \mathtt{C}_n \le 1. \tag{4.64}$$

By Equation (A.65)

$$\frac{\Gamma_d\left(\frac{\beta+\lambda n+m_n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n}{2}\right)} < \left(\frac{\beta+\lambda n+m_n}{2}\right)^{dm_n/2} \quad \text{and} \quad \frac{\Gamma_d\left(\frac{\beta+\lambda n+m_n+M_n}{2}\right)}{\Gamma_d\left(\frac{\beta+\lambda n+M_n}{2}\right)} > \left(\frac{\beta+\lambda n+m_n+M_n-d}{2e}\right)^{dm_n/2}, \tag{4.65}$$

and hence, for sufficiently large $n$ (such that $M_n \ge d$)

$$\mathtt{B}_n \le \left(\frac{e(\beta+\lambda n+m_n)}{\beta+\lambda n+m_n+M_n-d}\right)^{dm_n/2} \le \left(\frac{e(\beta+\lambda+1)}{\lambda}\right)^{dm_n/2}. \tag{4.66}$$

Let

$$\begin{aligned}
A_n &= (\alpha+\lambda n)\Sigma_0 + \sum_{i\in I_{\max}} (x_i - \overline{\mathbf{x}_{I_{\max}}})(x_i - \overline{\mathbf{x}_{I_{\max}}})^\top \\
B_n &= \sum_{i\in I_{\min}} (x_i - \overline{\mathbf{x}_{I_{\min}\cup I_{\max}}})(x_i - \overline{\mathbf{x}_{I_{\min}\cup I_{\max}}})^\top \\
C_n &= \frac{\kappa_0(m_n+M_n)}{\kappa_0+m_n+M_n}(\overline{\boldsymbol{x}_{I_{\min}\cup I_{\max}}} - \mu_0)(\overline{\boldsymbol{x}_{I_{\min}\cup I_{\max}}} - \mu_0)^\top \\
D_n &= \sum_{i\in I_{\max}} (x_i - \overline{\mathbf{x}_{I_{\min}\cup I_{\max}}})(x_i - \overline{\mathbf{x}_{I_{\min}\cup I_{\max}}})^\top - \sum_{i\in I_{\max}} (x_i - \overline{\mathbf{x}_{I_{\max}}})(x_i - \overline{\mathbf{x}_{I_{\max}}})^\top
\end{aligned} \tag{4.67}$$

Then

$$\Sigma(\boldsymbol{x}_{I_{\max}}) \succeq A_n, \quad \Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}}) = A_n + B_n + C_n + D_n. \tag{4.68}$$

Hence, by Lemma 4.4

$$\frac{|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})|}{|\Sigma(\boldsymbol{x}_{I_{\max}})|} \le \frac{|A_n+B_n+C_n+D_n|}{|A_n|} = \left|I_d + (B_n+C_n+D_n)A_n^{-1}\right| \le e^{\operatorname{tr}\left((B_n+C_n+D_n)A_n^{-1}\right)} \tag{4.69}$$

so

$$\mathtt{E}_n \le e^{\operatorname{tr}\left((B_n+C_n+D_n)\left(\frac{A_n}{\beta+\lambda n+M_n}\right)^{-1}\right)} \tag{4.70}$$

We now have

$$\begin{aligned}
\mathrm{tr}(B_n) &= \mathrm{tr}\Big(\sum_{i\in I_{\min}} (x_i - \overline{\mathbf{x}_{I_{\min}}})(x_i - \overline{\mathbf{x}_{I_{\min}}})^\top\Big) = \sum_{i\in I_{\min}} \mathrm{tr}\big((x_i - \overline{\boldsymbol{x}_{I_{\min}}})(x_i - \overline{\boldsymbol{x}_{I_{\min}}})^\top\big) = \\
&= \sum_{i\in I_{\min}} \big\|x_i - \overline{\boldsymbol{x}_{I_{\min}}}\big\|^2 \le m_n(2r)^2 \\
\mathrm{tr}(C_n) &= \frac{\kappa_0(m_n+M_n)}{\kappa_0+m_n+M_n}\|\overline{\boldsymbol{x}_{I_{\min}\cup I_{\max}}} - \mu_0\|^2 \le \kappa_0(2r)^2.
\end{aligned} \tag{4.71}$$

Moreover it is straightforward to compute that

$$D_n = M_n(\overline{\boldsymbol{x}_{I_{\max}\cup I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}})(\overline{\boldsymbol{x}_{I_{\max}\cup I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}})^\top \tag{4.72}$$

and

$$\begin{aligned}
\overline{\boldsymbol{x}_{I_{\max}\cup I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}} &= \frac{m_n}{m_n+M_n}\overline{\boldsymbol{x}_{I_{\min}}} + \frac{M_n}{m_n+M_n}\overline{\boldsymbol{x}_{I_{\max}}} - \overline{\boldsymbol{x}_{I_{\max}}} = \\
&= \frac{m_n}{m_n+M_n}(\overline{\boldsymbol{x}_{I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}})
\end{aligned} \tag{4.73}$$

and hence

$$\mathrm{tr}(D_n) \le \frac{M_n m_n^2}{(m_n+M_n)^2}\mathrm{tr}\big((\overline{\boldsymbol{x}_{I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}})(\overline{\boldsymbol{x}_{I_{\min}}} - \overline{\boldsymbol{x}_{I_{\max}}})^\top\big) \le m_n(2r)^2. \tag{4.74}$$

By (4.71) and (4.74)

$$\mathrm{tr}(B_n + C_n + D_n) \le (2m_n + \kappa_0)(2r)^2. \tag{4.75}$$

Moreover (using a natural upper bound $M_n \le n$)

$$\frac{A_n}{\beta+\lambda n+M_n} \succeq \frac{\alpha+\lambda n}{\beta+\lambda n+M_n}\Sigma_0 \succeq \frac{\lambda}{\beta+\lambda+1}\Sigma_0 \tag{4.76}$$

Thus

$$\mathrm{tr}\left(\Big(\frac{A_n}{\beta+\lambda n+M_n}\Big)^{-1}\right) \le \frac{\beta+\lambda+1}{\lambda}\mathrm{tr}(\Sigma_0^{-1}) \tag{4.77}$$

By Lemma A.10 (part (a)), (4.75) and (4.77)

$$\mathrm{tr}\Big((B_n + C_n + D_n)\left(\frac{A_n}{\beta+\lambda n+M_n}\right)^{-1}\Big) \le (2m_n+\kappa_0)(2r)^2\frac{\beta+\lambda+1}{\lambda}\mathrm{tr}(\Sigma_0^{-1}). \tag{4.78}$$

Equations (4.70) and (4.78) give

$$\log \mathrm{E}_n \le (2m_n+\kappa_0)(2r)^2\frac{\beta+\lambda+1}{\lambda}\mathrm{tr}(\Sigma_0^{-1}) \tag{4.79}$$

Note that by Lemma A.8

$$|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})| \le \big\|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})\big\|^d \tag{4.80}$$

and, using the triangle inequality and Lemma A.9

$$\begin{aligned}
\big\|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})\big\| &\le (\alpha+\lambda n)\|\Sigma_0\| + (m_n+M_n)(2r)^2 + \frac{\kappa_0}{\kappa_0+m_n+M_n}(2r)^2 \le \\
&\le (\alpha+\lambda n)\|\Sigma_0\| + 4(m_n+M_n+1)r^2.
\end{aligned} \tag{4.81}$$

Using (4.81) and (4.80) we get

$$\begin{aligned}\frac{|\Sigma(\boldsymbol{x}_{I_{\min}\cup I_{\max}})|}{|\Sigma(\boldsymbol{x}_{I_{\min}})|} &\leq \frac{\big((\alpha+\lambda n)\|\Sigma_0\|+4(m_n+M_n+1)r^2\big)^d}{|(\alpha+\lambda n)\Sigma_0|} = \\ &= \frac{\big(\|\Sigma_0\|+4\frac{m_n+M_n+1}{\alpha+\lambda n}r^2\big)^d}{|\Sigma_0|} \leq \frac{\big(\|\Sigma_0\|+\frac{8}{\lambda}r^2\big)^d}{|\Sigma_0|}.\end{aligned} \tag{4.82}$$

Putting together (4.64), (4.66), (4.79) and (4.82) and plugging it to (4.62) we obtain that for sufficiently large $n$

$$\frac{g_{aNIW,n}(\boldsymbol{u}_{1:n}\,|\,\hat{\mathcal{I}}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n}\,|\,\tilde{\mathcal{I}}_n)} \leq e^{\kappa_0(2r)^2\frac{\beta+\lambda+1}{\lambda}\mathrm{tr}(\Sigma_0^{-1})}\left[\frac{\big(e(\beta+\lambda+1)\big(\|\Sigma_0\|+\frac{8}{\lambda}r^2\big)\big)^{d/2}}{\lambda^{d/2}|\Sigma_0|^{1/2}}e^{2(2r)^2\frac{\beta+\lambda+1}{\lambda}\mathrm{tr}(\Sigma_0^{-1})}\right]^{m_n}. \tag{4.83}$$

It is easily seen that

$$\frac{\mathcal{P}_{\pi,n}(\hat{\mathcal{I}}_n)}{\mathcal{P}_{\pi,n}(\tilde{\mathcal{I}}_n)} = \alpha\frac{(m_n-1)!(M_n-1)!}{(m_n+M_n-1)!} \leq \alpha\frac{m_n!(M_n-1)!}{(m_n+M_n-1)!} \leq \alpha\left(\frac{m_n}{M_n-1}\right)^{m_n} \tag{4.84}$$

As $\liminf_{n\to\infty}\frac{m_n}{M_n}=0$, then taking into account (4.83) and (4.84) we get

$$\liminf_{n\to\infty}\frac{g_{aNIW,n}(\boldsymbol{u}_{1:n}\,|\,\hat{\mathcal{I}}_n)}{g_{aNIW,n}(\boldsymbol{u}_{1:n}\,|\,\tilde{\mathcal{I}}_n)}\frac{\mathcal{P}_{\pi,n}(\hat{\mathcal{I}}_n)}{\mathcal{P}_{\pi,n}(\tilde{\mathcal{I}}_n)} = 0, \tag{4.85}$$

which contradicts the definition of the MAP partition. This is a contradiction and hence the result is established.

## 4.2 From asymptotic formulae to score functions

In Section 2.2 we derived a score function $\Delta^{aNIW}_{P,\lambda}$ for partitions of the observation space, based on asymptotics of the posterior probability of partitions. The prior distribution over the within-cluster covariance contains a parameter $\Sigma_0$, which is the expected within-cluster covariance. There is another parameter $\lambda$; the degrees of freedom of the inverse Wishart distribution are $\nu_0=\alpha+\lambda n$, where $\alpha$ is a non-negative constant and $\lambda>0$ is a constant. The parameter $\lambda$ gives the strength of dependence on $\Sigma_0$; for small $\lambda$ the within-cluster covariance structure can vary substantially from $\Sigma_0$, while for $\lambda=\infty$, each within-cluster covariance structure is exactly $\Sigma_0$.

By the *score function* we mean a function that can be used to assess a clustering proposal. For example, between-cluster variance is a reasonable score function, which is maximised by the $k$-means algorithm. In the context of our research, a natural idea for a score function is an asymptotic formula $\Delta^{aNIW}_{P,\lambda}$, where $P$ is the distributional limit of empirical probability distributions $\hat{P}_n=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}$, where $X_1,\ldots,X_n\overset{\text{iid}}{\sim}P$. We approximate $\Delta^{aNIW}_{P,\lambda}(\mathcal{A})$ by $\Delta^{aNIW}_{\hat{P}_n,\lambda}(\mathcal{A})$. Note that $\Delta^{aNIW}_{\hat{P}_n,\lambda}(\mathcal{A})$ depends on the sample only through the estimates of set probabilities $\hat{P}_n(A)=\frac{1}{n}|\{i\colon X_i\in A\}|$ and estimates of the conditional covariance structures

$$\mathbf{V}_{\hat{P}_n}(X\,|\,X\in A) = \frac{1}{n}\sum_{i\colon X_i\in A}X_iX_i^\top - \left(\frac{1}{n}\sum_{i\colon X_i\in A}X_i\right)\left(\frac{1}{n}\sum_{i\colon X_i\in A}X_i\right)^\top. \tag{4.86}$$

Therefore, $\Delta_{\widehat{P}_n,\lambda}$ approximates $\Delta_{P,\lambda}$ well if and only if $\hat{P}_n(A)$ approximates $P(A)$ well for each $A \in \mathcal{A}$ and $\mathbf{V}_{\hat{P}_n}(X \mid X \in A)$ approximates $\mathbf{V}_P(X \mid X \in A)$ well for each $A \in \mathcal{A}$.

Let $\mathcal{J}$ be a partition of indices from 1 to $n$ which defines the clustering $\big\{\{x_i \colon i \in J\} \colon J \in \mathcal{J}\big\}$ of the data $x_1, \ldots, x_n$. The *score* for this clustering is:

$$\begin{aligned}\hat{\Delta}_\lambda(\mathcal{J}) = &-\frac{1}{2}\sum_{J\in\mathcal{J}}\big(\hat{P}_n(\boldsymbol{x}_J)+\lambda\big)\log\Big|\frac{\lambda}{\hat{P}_n(\boldsymbol{x}_J)+\lambda}I_d + \frac{\hat{P}_n(\boldsymbol{x}_J)}{\hat{P}_n(\boldsymbol{x}_J)+\lambda}\Sigma_0^{-1}\hat{\mathbf{V}}_n(\boldsymbol{x}_J)\Big| + \\ &+\sum_{J\in\mathcal{J}}\hat{P}_n(\boldsymbol{x}_J)\log\hat{P}_n(\boldsymbol{x}_J).\end{aligned} \tag{4.87}$$

where

$$\hat{P}_n(\boldsymbol{x}_J) = \frac{|J|}{n}, \quad \hat{\mathbf{V}}_n(\boldsymbol{x}_J) = \frac{1}{n}\sum_{i\in J}(x_i - \overline{\boldsymbol{x}_J})(x_i - \overline{\boldsymbol{x}_J})^\top. \tag{4.88}$$

The *score* formula (4.87) is obtained from $\Delta_{\hat{P}_n,\lambda}(\cdot)$ by dropping the additive constant $-\frac{1}{2}\log|\Sigma_0| - \frac{d}{2}$.

In the $\lambda = 0$ limit, the objective is:

$$\hat{\Delta}_0(\mathcal{J}) = -\frac{1}{2}\sum_{J\in\mathcal{J}}\hat{P}_n(\boldsymbol{x}_J)\log|\hat{\mathbf{V}}_n(\boldsymbol{x}_J)| + \sum_{J\in\mathcal{J}}\hat{P}(\boldsymbol{x}_J)\log\hat{P}(\boldsymbol{x}_J) + \frac{1}{2}\log|\Sigma_0|. \tag{4.89}$$

Note that, no matter what the within-cluster variance actually is, if $|J| < d$ for some $J \in \mathcal{J}$ then $|\hat{\mathbf{V}}_n(\boldsymbol{x}_J)| = 0$ and hence $\hat{\Delta}_0(\mathcal{J}) = \infty$.

The results of applying this strategy to score the cluster proposals of simulated datasets are contained in Chapter 5. We now discuss some considerations which give a suitable lower bound on the number of data points needed per cluster so that (4.87) gives a reasonable estimate of (4.3). The worst case scenario is $\lambda = 0$.

Conditioned on belonging to the same cluster, observations are i.i.d. normal. Let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, C)$, where $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}_+^{d\times d}$. Let $\hat{C}$ be the maximum likelihood estimator of the covariance matrix, i.e.

$$\hat{C} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^\top. \tag{4.90}$$

According to Goodman (1963), for $n \geq d$:

$$\frac{n^d}{\det C}\det\hat{C} \sim \prod_{i=1}^d Z_i, \tag{4.91}$$

where $Z_1, \ldots, Z_d$ are independent random variables and $Z_i \sim \chi^2_{n-i+1}$. Therefore

$$\log\det\hat{C} \sim \log\det C - d\log n + \sum_{i=1}^d \log Z_i. \tag{4.92}$$

If $Y \sim \mathrm{Gamma}(\alpha, \beta)$ then $\mathbb{E}\log Y = \psi(\alpha) - \log\beta$ and $\mathrm{Var}\log Y = \psi_1(\alpha)$, where $\psi$ and $\psi_1$ are digamma and trigamma functions respectively ($\psi(\alpha) = \Gamma'(\alpha)$, $\psi_1(\alpha) = \psi'(\alpha)$). As $Z_i \sim \mathrm{Gamma}(\frac{n-i+1}{2}, \frac{1}{2})$, we get

$$\mathbb{E}\log Z_i = \psi\left(\frac{n-i+1}{2}\right) - \log\frac{1}{2}, \quad \mathrm{Var}\log Z_i = \psi_1\left(\frac{n-i+1}{2}\right). \tag{4.93}$$

Putting (4.92) and (4.93) together, we get

$$\begin{aligned}\mathbb{E}\log\det\hat{C} &= \log\det C + \sum_{i=1}^{d}\left(\psi\left(\frac{n-i+1}{2}\right) - \log\frac{n}{2}\right)\\ \operatorname{Var}\log\det\hat{C} &= \sum_{i=1}^{d}\psi_1\left(\frac{n-i+1}{2}\right)\end{aligned} \tag{4.94}$$

By Alzer (1997) for $x > 0$ we have $|\psi(x) - \log x| < \frac{1}{2x}$ and hence for $i = 1, 2, \ldots, d$

$$\begin{aligned}\left|\psi\left(\frac{n-i+1}{2}\right) - \log\frac{n}{2}\right| &< \left|\psi\left(\frac{n-i+1}{2}\right) - \log\frac{n-i+1}{2}\right| + \left|\log\frac{n-i+1}{2} - \log\frac{n}{2}\right| <\\ &< \frac{1}{n-i+1} + \log\left(1 + \frac{i-1}{n-i+1}\right) < \frac{i}{n-i+1}.\end{aligned} \tag{4.95}$$

This implies that

$$\left|\mathbb{E}\log\det\hat{C} - \log\det C\right| \leq \frac{d^2}{n-d+1}. \tag{4.96}$$

Due to Guo and Qi (2013) we have $\psi_1(x) < e^{\frac{1}{x}} - 1$ and so

$$\operatorname{Var}\log\det\hat{C} < d\left(e^{\frac{1}{n-d+1}} - 1\right). \tag{4.97}$$

The bias $\tau_{n,d} = \sum_{i=1}^{d}\left(\psi\left(\frac{n-i+1}{2}\right) - \log\frac{n}{2}\right)$ is relatively large for small sample sizes. Taking these considerations into account (especially (4.96)) we propose a rule of thumb in which the approximation should be applied is where the number of observations in every cluster is of order $d^2$.

## 4.3 Summary and Perspectives for Future Work

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 were 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). We also computed the asymptotic limit of the posterior of an induced partition (Proposition 4.1) and we established some properties of the limit (Lemma 4.4 and Proposition 4.5). Finally, in Section 4.2 we suggest how to use the empirical of the limit to score cluster proposals (the experimental analysis of this approach is contained in Chapter 5).

In future work we intend to establish an analogue of Theorem 3.21 for this adjusted Normal-Inverse-Wishart model. Moreover we plan to investigate further the clustering properties of the $\Delta_{P,\lambda}^{aNIW}$ function (and its empirical equivalent). An interesting starting point is the property of the $\Delta_P^{NIW}$ function when $P$ is the uniform distribution on $[0, 1]$. As we showed towards the end of Chapter 2, every partition of $[0, 1]$ into subsegments maximises the

$\Delta_P^{NIW}$ function. We plan to show that for the $\Delta_{P,\lambda}^{aNIW}$ function, the maximiser is unique (just like with the $\Delta_P^{NN}$ function, cf. Proposition 2.30) and to investigate the limit as $\lambda \to 0$. This may be thought of analogously to the 'viscosity solution' that is often popular in partial differential equations; a p.d.e. may have many solutions when the viscosity parameter is 0 in situations where there is uniqueness for positive viscosity. The solution to the inviscid equation which is of interest often corresponds to solving the equation with viscosity and letting the viscosity parameter tend to 0. Letting lambda go to zero in the problem defined by (4.1) sometimes gives a particular solution of the problem where the prior is a traditional prior, which does not depend on $n$. By choosing $\Sigma_0$ as the total covariance matrix, this should give the solution with the smallest number of clusters that solves the $\lambda = 0$ problem.

# Chapter 5

# Experimental Results

In Chapter 4 we derived a score function and indicated the quantity of data required to ensure that the 'empirical' score function approximates well the limit of log posterior. We now give some examples to show the clusterings favoured by this score function.

Of course, the problem of finding the clustering which maximised the score function is computationally far too demanding even with relatively modest examples, so we derive candidate clusterings by other means and then see which are favoured by the score function.

In our examples, we will propose clusterings using the $K$-means algorithm and apply the score function, to see (a) which value of $K$ gives the greatest score and (b) whether the 'true' clustering gives a greater score than the best $K$-means clustering.

We have two hyperparameters to consider; $\Sigma_0$, which is the a-priori expected within-cluster covariance mean and which gives the strength of the prior assessment. The other hyperparameter is $\lambda$ that controls the strength of the prior information on the within-cluster covariance.

In practise, we find that the *total covariance matrix* of the entire data set gives a good value for $\Sigma_0$. This prevents the score function from favouring too many clusters; if the parameter $\lambda$ is sufficiently small, neither is the number of clusters underestimated. The single-cluster solution gives a good reference point; with this choice of $\Sigma_0$, the score function for a single cluster is 0. Indeed, let $[n] = \{1, \ldots, n\}$ then:

$$\begin{aligned}\hat{\Delta}_{P,\lambda}(\{[n]\}) &= -\frac{1}{2}\big(\hat{P}(\boldsymbol{x}_{[n]}) + \lambda\big) \log \Big| \frac{\lambda}{\hat{P}(\boldsymbol{x}_{[n]}) + \lambda} I_d + \frac{\hat{P}(\boldsymbol{x}_{[n]})}{\hat{P}_n(\boldsymbol{x}_{[n]}) + \lambda} \hat{\mathbf{V}}_n(\boldsymbol{x}_{[n]})^{-1} \hat{\mathbf{V}}_n(\boldsymbol{x}_{[n]}) \Big| + \\ &+ \hat{P}_n(\boldsymbol{x}_{[n]}) \log \hat{P}_n(\boldsymbol{x}_{[n]}) = \\ &= -\frac{1}{2}\big(\hat{P}(\boldsymbol{x}_{[n]}) + \lambda\big) \log |I_d| + \hat{P}_n(\boldsymbol{x}_{[n]}) \log \hat{P}_n(\boldsymbol{x}_{[n]}) = 0,\end{aligned} \tag{5.1}$$

since $\hat{P}_n(\boldsymbol{x}_{[n]}) = 1$. Therefore, for different clusterings, it can easily be seen whether or not the clustering gives a better or worse score than for the single-cluster solution.

## 5.1 Metrics for Clustering

For two proposed clusterings, it is important to be able to decide whether they are close to each other or whether they differ markedly. It is useful to have measures of divergence between clusterings. We consider two metrics, presented in Wade et al. (2018): if $\mathcal{J}_1$ and $\mathcal{J}_2$ are two partitions of the set $[n]$ then

- (Binder loss) $f_B(x) = x^2$ and

$$d_B(\mathcal{J}_1, \mathcal{J}_2) = \sum_{J_1 \in \mathcal{J}_1} f_B(|J_1|/n) + \sum_{J_2 \in \mathcal{J}_2} f_B(|J_2|/n) - 2 \sum_{J_1 \in \mathcal{J}_1, J_2 \in \mathcal{J}_2} f_B(|J_1 \cap J_2|/n) \quad (5.2)$$

- (Variation of Information) $f_{VI}(x) = x \log x$ and

$$d_{VI}(\mathcal{J}_1, \mathcal{J}_2) = \sum_{J_1 \in \mathcal{J}_1} f_{VI}(|J_1|/n) + \sum_{J_2 \in \mathcal{J}_2} f_{VI}(|J_2|/n) - 2 \sum_{J_1 \in \mathcal{J}_1, J_2 \in \mathcal{J}_2} f_{VI}(|J_1 \cap J_2|/n) \quad (5.3)$$

Recall the mathematical definition of a metric; $d$ is a metric if $d \geq 0$, $d(A, B) = d(B, A)$, $d(A, B) \leq d(A, C) + d(B, C)$ for any $A, B, C$. The fact that both, the Binder loss and the Variation of Information, are metrics on the space of clusterings is listed in Wade et al. (2018) as Property 1. The proof of this property for the Variation of Information is contained in Meilă (2007) and the proof for the Binder loss, as the authors of Wade et al. (2018) points out, follows from the fact that this can be consider as a Hamming distance between the indicator representations of the clusterings.

Let $\mathbf{0}_n$ and $\mathbf{1}_n$ be the $n$-cluster and 1-cluster partitions of $[n]$ respectively. By Property 4 in Wade et al. (2018) we have that for any partitions $\mathcal{J}_1$ and $\mathcal{J}_2$ of $[n]$

$$d_B(\mathcal{J}_1, \mathcal{J}_2) \leq d_B(\mathbf{0}_n, \mathbf{1}_n) = 1 - \frac{1}{n} \quad \text{and} \quad d_{VI}(\mathcal{J}_1, \mathcal{J}_2) \leq d_{VI}(\mathbf{0}_n, \mathbf{1}_n) = \log n \quad (5.4)$$

To put these distances on the same scale, we consider their normalised versions, i.e. $\bar{d}_B(\mathcal{J}_1, \mathcal{J}_2) = d_B(\mathcal{J}_1, \mathcal{J}_2)/d_B(\mathbf{0}_n, \mathbf{1}_n)$ and $\bar{d}_{VI}(\mathcal{J}_1, \mathcal{J}_2) = d_{VI}(\mathcal{J}_1, \mathcal{J}_2)/d_{VI}(\mathbf{0}_n, \mathbf{1}_n)$.

We now illustrate the performance of the score function by considering several data sets.

## 5.2 Example: Simulated mixtures of Gaussians

For our first example, we generate $n = 10^4$ data points according to the generation scheme for cluster assignment, cluster distribution and within-cluster observation, which we modify by truncating the stick-breaking construction at some small level $K = 5$ (so that 5 is the true number of clusters). We therefore have the true underlying distribution; the within-cluster distribution, the generating mechanism for each cluster and the theoretical proportions for each cluster. We know the 'true' cluster assignment.

For various values of $K$, $K = 1, \ldots, 12$, we run the $K$-means algorithm (Steinhaus (1956), MacQueen et al. (1967)) to obtain clusterings (Figure 5.1) and we score them using our score function $\hat{\Delta}_\lambda$, for a selection of 15 $\lambda$ values, which include $\lambda = 0$ and $\lambda = \infty$. The results are presented on Figure 5.2.

Note that all the plots start from the origin, since $\hat{\Delta}_\lambda(\{1, \ldots, n\}) = 0$ by (5.1). Also, other than for the single-cluster solution, the respective plots do not coincide, since Proposition 4.3 (a) establishes that $\Delta_{\lambda,P}(\mathcal{A})$ is *decreasing* in $\lambda$ for any fixed $\mathcal{A}$.

When we do not place constraints on the minimum acceptable cluster size, the number of clusters is clearly overestimated for small $\lambda$.

As the value of $\lambda$ increases, the score converges to the $\hat{\Delta}_\infty$ function, which is the empirical equivalent of the Delta function that was analysed in Chapter 3, in this case the clusters tend to adjust themselves to the covariance matrix $\Sigma_0$. Since we have set $\Sigma_0$ equal to the total covariance matrix, it is not surprising that $\hat{\Delta}_\infty$ favours the partition with just one cluster.

Recall that Figure 5.2 gives the scores of the clusterings produced by the $K$-means algorithm for different values of $K$. The true value of $K$ is 5, yet the $K$-means clustering into 5 clusters performs badly; for a large range of values for $\lambda$, the $\Delta_\lambda$ scores shown on Figure 5.2 gives a local maximum for 3 clusters and, would suggest either the 8-mean clustering or the 3-mean clustering produced by $K$-means.

Since we know the true clustering structure in this case, we can compute the normalised distances between the various $K$-means solutions and the true clustering. The 8-means solution was closest to the true partition according to both metrics (Binder loss and Variation of Information). The information is given in the plot (Figure 5.3) in the shaded area. Note that the normalised distance values are transformed by a decreasing function $F$ and hence the larger outputs correspond to 'better' partition, so that it corresponds to the behaviour of $\Delta_\lambda$ function. The details of the transformation are given in the caption of Figure 5.3.

Another informative part of Figure 5.3 are black segments that represent the relation between the maximal score of the $\hat{\Delta}_\lambda$ function among $K$-means proposals for different values of $\lambda$ and the $\hat{\Delta}_\lambda$ of the true clustering. It should be noted that for $\lambda < .5$ the true partition is scored higher than the best $K$-means proposal.

This indicates two things: (a) the $\hat{\Delta}$ score function gives higher scores to 'correct' partitions and (b) in this example, the K-means algorithm is not performing very well; none of the partitions chosen by $K$-means were close to the true partition.

In this example, there are well-defined clusters, but $K$-means has not found them. If these clusterings are scored using the formula then the 'true' clustering scores more highly than any of the $K$-means clusterings.

Figure 5.4 presents another representative example (regardless of the data dimension) of

the behaviour of the $\hat{\Delta}_\lambda$ score function for mixture of Gaussians and illustrates the following properties:

- If the true clustering is not detected by the $K$-means algorithm (which is often the case when the covariance structure varies between clusters and when a within-cluster covariance $\Lambda$ exhibits substantial correlation - i.e. $\left|\frac{\Lambda_{ij}}{\sqrt{|\Lambda_{ii}\Lambda_{ij}|}}\right|$ is close to 1 for some $i, j$) then the true clustering has a higher score than any of the $K$-means clusterings, at least for $\lambda \in [0, \bar{\lambda}]$ for some $\bar{\lambda} < 1$.
- If the true clustering is not detected by the $K$-means algorithm then the maximisers of $\hat{\Delta}_0$ score do not always correspond to the $K$-means clustering proposal closest to the true clustering according to some metric. On the other hand, usually there is a range $[\lambda_1, \lambda_2]$ such that for $\lambda$ within this range the partition selected as maximiser for $\hat{\Delta}_\lambda$ is the proposal which is closest to the true clustering.
- The number of clusters in the maximiser of the $\hat{\Delta}_\lambda$ function tends to be a non-increasing function of $\lambda$. However, this is not always the case, as shown on Figure 5.5.

## 5.3 Example: The Fisher Iris Data

In this section we investigate the performance of the $\hat{\Delta}_\lambda$ function on the standard, 4-dimensional `iris` dataset. For a reminder, the dataset contains information about 3 species of irises, 50 observation each. We use the suggestions of the $K$-means algorithm for $K = 1, 2, \ldots, 10$ and score them using $\hat{\Delta}_\lambda$ function for a selection of 50 values of the $\lambda$ parameter. We include this example to show that, as the theory suggests, our method is not well suited for a relatively small sample sizes. The cluster sizes in the 10-means clustering are 23, 22, 19, 18, 17, 16, 12, 11, 8 and 4. Clearly the within group covariance structure in 4-dimensional space cannot be estimated with samples of this size. We should not expect good results under such circumstances, as indicated by the considerations in Section 4.2.

The results are presented in Figure 5.6 (see Section 5.2 for the detailed description of the diagnostic plot) and they support these predictions. There is a significant instability in the behaviour of the score function, just as the theory suggests.

We do not have access to a larger database concerning irises. Instead we decide to mimic the original sample by generating a mixture of Gaussians. We analyse a realisation of size $10^4$ of the mixture of three 4-dimensional Gaussian distribution, with the same mean-covariance structure as the original dataset. The results are shown on Figure 5.7. Here the $\hat{\Delta}_\lambda$ score function nicely detects the best clustering proposal suggested by $K$-means ($K = 3$) and its value is lower than the score of the true clustering.

This example confirms that the $\hat{\Delta}_\lambda$ score should not be used when there is not enough information to estimate the covariance structure of the clusters.

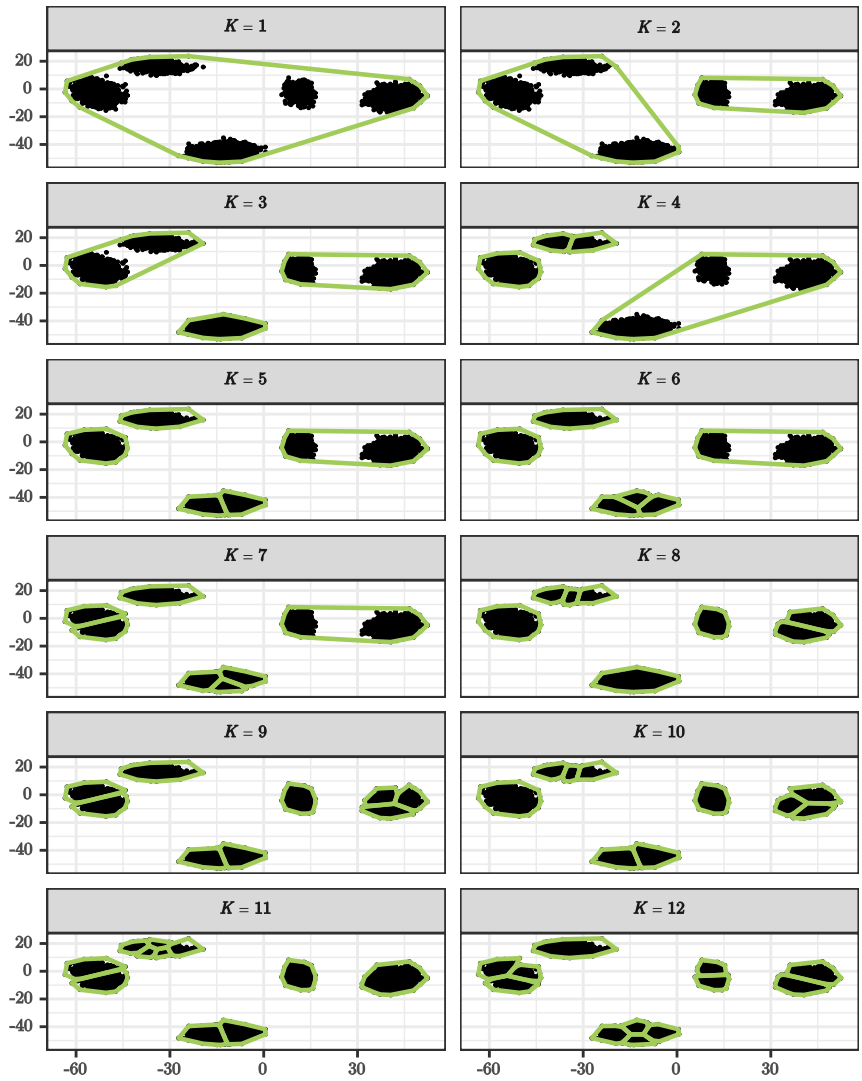



Figure 5.1: The $K$-means clustering proposals for the mixture of Gaussians example. The boundary of convex hulls of clusters are shown in green.

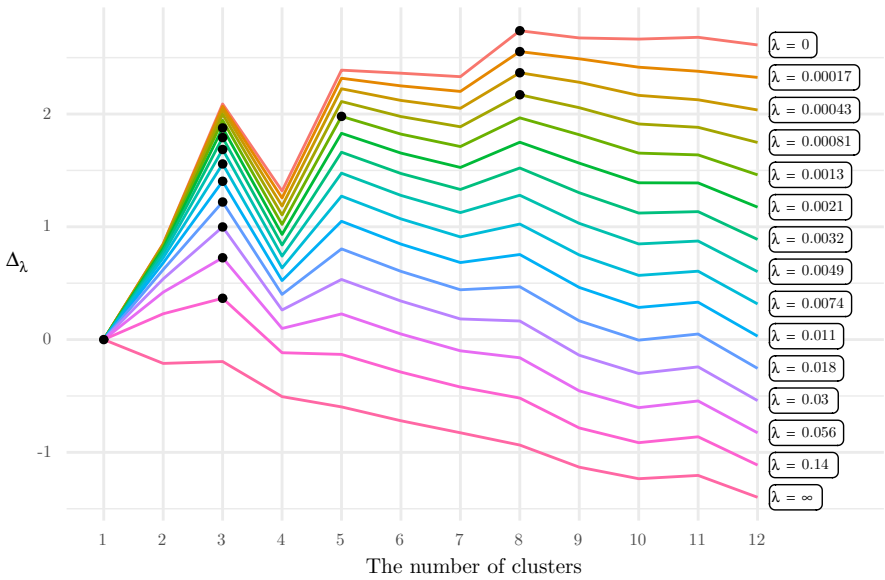



Figure 5.2: The $\hat{\Delta}_\lambda$ scores of the mixture of Gaussians example. The maximal values of different $\hat{\Delta}_\lambda$ curves for different $\lambda$ values are denoted by black points. The $\lambda$ parameters were chosen so that the values of $\hat{\Delta}_\lambda$ for the 12-means clustering proposal are evenly spread.

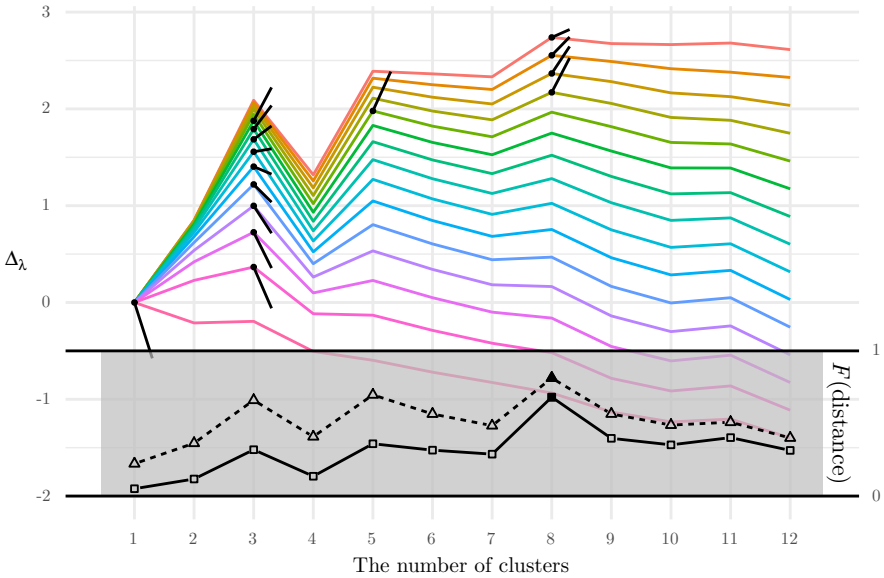



Figure 5.3: Diagnostic plot relevant to Figure 5.2. As before, the maximal values of different $\hat{\Delta}_\lambda$ curves for different $\lambda$ values are denoted by black points. Black segments beginning with these black points end with the $\hat{\Delta}_\lambda$ value of the true clustering for relevant $\lambda$. This means that whenever those segments have a positive slope, the true clustering is scored higher that any of the $K$-means proposal. Two black plots on the grey background represent $F(\tilde{b}_k)$ (squares) and $F(\tilde{v}_k)$ (triangles), where $\tilde{b}_k$ and $\tilde{v}_k$ represent the normalised Binder or VoI distances of the $k$-means proposal to the true clustering and $F(x) = \frac{.04}{.01+x}$ is a decreasing transformation. The formula of the transformation was chosen so that the results fit nicely into $[0, 1]$ interval (note the secondary $y$-axis labels on the right). Simply using $F(x) = 1 - x$ was not a good idea since relatively small distance values are quite common and difficult to distinguish visually. The points representing maximal values are painted black and respective $K$-means proposals are the closest (according to relevant distance measures) to the true clustering.

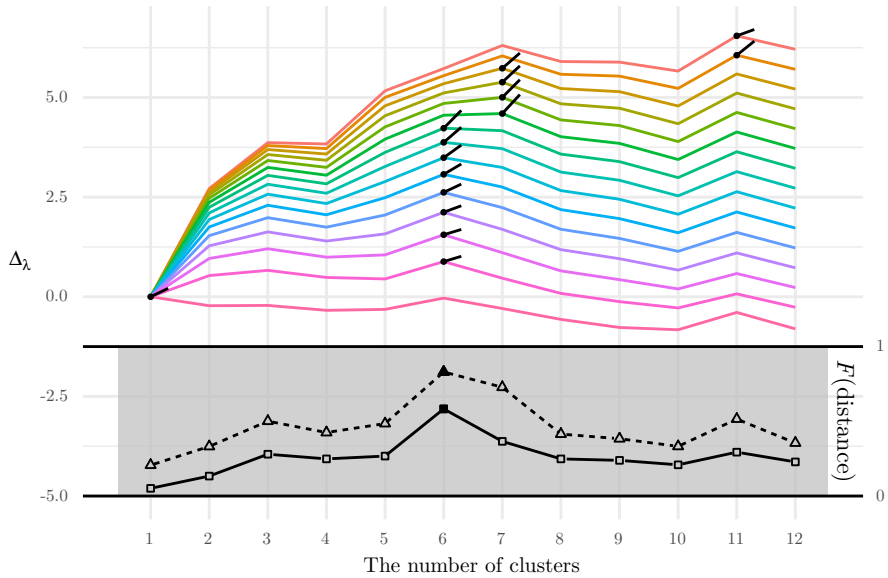



Figure 5.4: A representative example of the behaviour for the $\hat{\Delta}_\lambda$ score function. The data dimension is 4, the number of observations is $10^4$ and the cluster sizes in the true clustering were 3046, 1914, 297, 1555, 311, 1051 and 1826. Here $F(x) = \frac{.04}{.01+x}$.

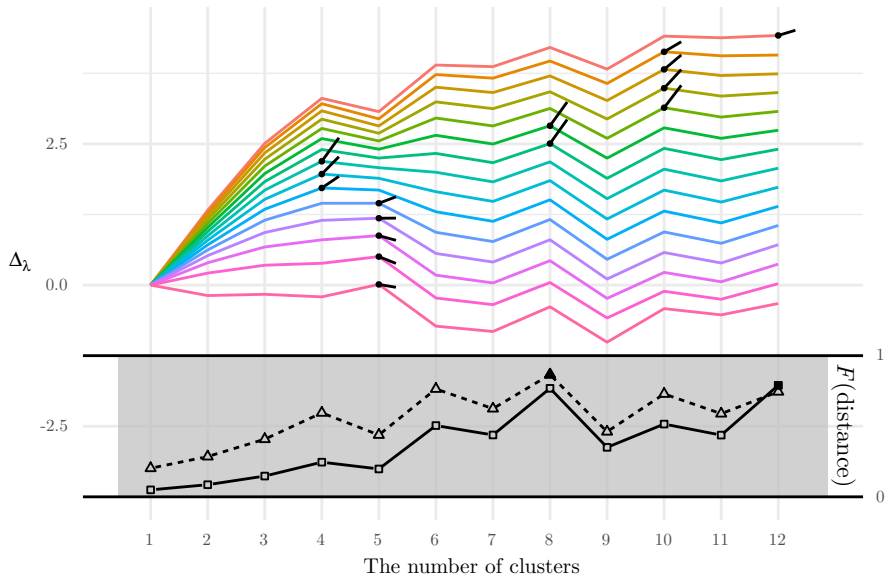



Figure 5.5: An example of the number of clusters of the maximiser of the $\hat{\Delta}_\lambda$ function among the $K$-means proposal not being a non-decreasing function of $\lambda$. Here $F(x) = \frac{.04}{.01+x}$.

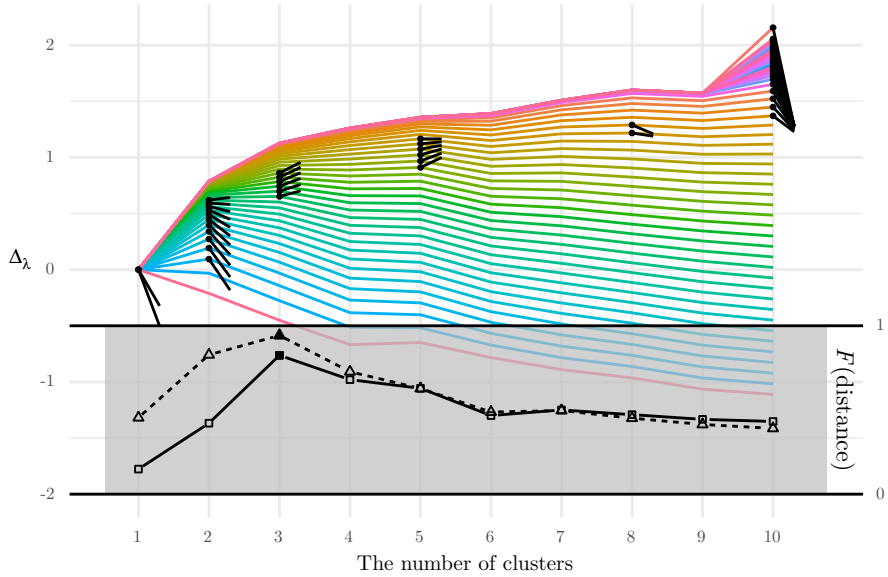



Figure 5.6: The diagnostic plot for the `iris` dataset. Here $F(x) = \frac{.01}{.001+x}$.

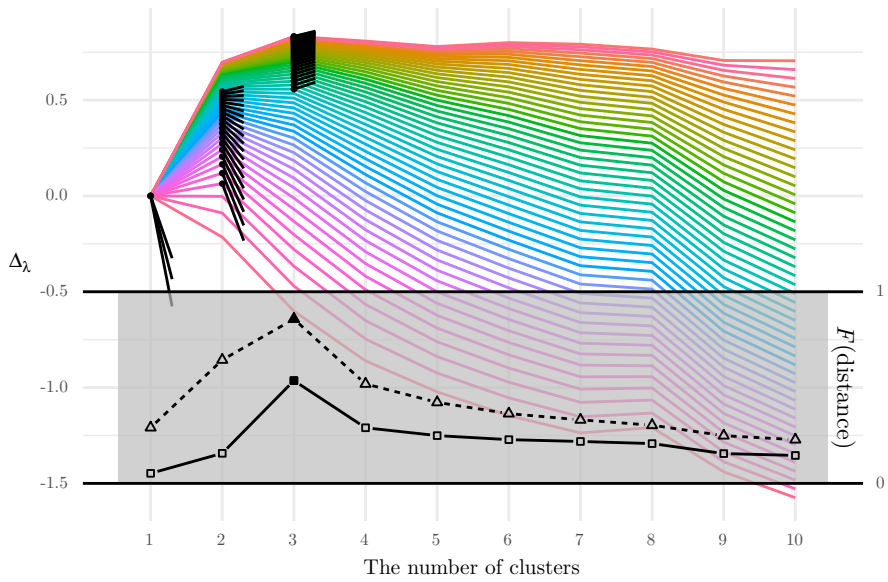



Figure 5.7: The diagnostic plot for the artificial mixture of Gaussians that mimics the mean-covariance structure of the `iris` dataset. Here $F(x) = \frac{.035}{.001+x}$.

# Appendix A

# Auxiliary Results

We start by fixing some notation.

**Notation.** If $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are real sequences, we write

- $a_n \simeq b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$.
- $a_n \asymp b_n$ if $0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$.
- $a_n \preceq b_n$ if $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$.
- $a_n = o(b_n)$, if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

Analogous notation is used when $a, b\colon \mathbb{R}_+ \to \mathbb{R}$ are real functions.

- We denote the convex hull of $\Delta^K$ and the origin by

$$\blacktriangle^K = \{(p_1, \ldots, p_K)\colon \sum_{k=1}^{K} p_k \leq 1, \forall_{k\leq K} p_k \in [0,1]\}. \tag{A.1}$$

- If $x \in \mathbb{R}^d$, by $\|x\|$ we mean the standard Euclidean norm. If $\Sigma \in \mathbb{R}^{d\times d}$ is a matrix, then $\|\Sigma\|$ is the operator norm, i.e.

$$\|\Sigma\| = \sup_{x\in\mathbb{R}^d\setminus 0} \|\Sigma x\|/\|x\|. \tag{A.2}$$

**Lemma A.1.** *Let $(V_i)_{i=1}^{\infty}$ be a sequence of random variables such that $V_i \sim \mathrm{Beta}(1-\beta, \alpha + i\beta)$ for some $\alpha > 0$ and $0 \leq \beta < 1$. Then $\sum_{i=1}^{\infty} \log(1 - V_i) \overset{a.s.}{=} -\infty$.*

*Proof.* If $\beta = 0$ then $V_i \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$. Since $\mathbb{P}(V_i < \frac{1}{2}) > 0$, by the Borel-Cantelli lemma we deduce that almost surely $\log(1 - V_i) < -\log 2$ for infinitely many indices $i$ and hence $\sum_{i=1}^{\infty} \log(1 - V_i) \overset{\text{a.s.}}{=} -\infty$.

For $\beta > 0$ note that by the concavity of the logarithm and the Jensen's inequality we get

$$\mathbb{E}\, \log(1 - V_i) \leq \log \mathbb{E}\,(1 - V_i) = \log\Big(1 - \frac{1-\beta}{\alpha + (i-1)\beta + 1}\Big). \tag{A.3}$$

Since $\log\left(1-\frac{1-\beta}{\alpha+(i-1)\beta+1}\right) \simeq -\frac{1-\beta}{\alpha+(i-1)\beta+1}$ and the series $\sum_{i=1}^{\infty}\frac{1-\beta}{\alpha+(i-1)\beta+1}$ is divergent, then by the limit comparison test the series $\sum_{i=1}^{\infty}\log\left(1-\frac{1-\beta}{\alpha+(i-1)\beta+1}\right)$ is also divergent and therefore by (A.3) and the linearity of the expected value

$$\mathbb{E}\sum_{i=1}^{\infty}\log(1-V_i) = -\infty. \tag{A.4}$$

We now prove that for $X \sim \mathrm{Beta}(a,b)$ then

$$\mathrm{Var}\log X = \psi_1(a) - \psi_1(a+b), \tag{A.5}$$

where $\psi_1$ is the *trigamma* function, defined by $\psi_1(z) = (\log\Gamma(z))''$. Let $f_{a,b}(x) = \frac{1}{\mathcal{B}(a,b)}x^{a-1}(1-x)^{b-1}$ be the density of the $\mathrm{Beta}(a,b)$ distribution. Consider the parameter $b$ as fixed and recall the definition of the *Fisher information value* for a single-parameter family of distributions:

$$\mathcal{I}(a) = \left(\mathbb{E}\frac{\partial}{\partial a}\log f_{a,b}(X)\right)^2 \tag{A.6}$$

It is a standard property of the Fisher information value (Bickel and Doksum, 2015, Proposition 3.4.4) that

$$\mathrm{Var}\frac{\partial}{\partial a}\log f_{a,b}(X) = \mathcal{I}(a) = -\mathbb{E}\frac{\partial^2}{\partial a^2}\log f_{a,b}(X), \tag{A.7}$$

Computing the exact formulas of the right- and left-hand sides of (A.7), we obtain

$$\mathrm{Var}\frac{\partial}{\partial a}\log f_{a,b}(X) = \mathrm{Var}\log X \quad \text{and} \quad \mathbb{E}\frac{\partial^2}{\partial a^2}\log f_{a,b}(X) = \frac{\partial^2}{\partial a^2}\frac{1}{\mathcal{B}(a,b)}. \tag{A.8}$$

Since $\mathcal{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, we easily obtain that $\frac{\partial^2}{\partial a^2}\frac{1}{\mathcal{B}(a,b)} = \psi_1(a+b) - \psi_1(a)$. Now (A.5) follows from (A.7) and (A.8).

By Abramowitz and Stegun, Formula 6.4.12 we get $\psi_1(z) = \frac{1}{z} + \frac{1}{2z^2} + o(\frac{1}{z^2})$. Hence by (A.5)

$$\mathrm{Var}\log(1-V_i) = \psi_1(\alpha+i\beta) - \psi_1\big(\alpha+(i-1)\beta+1\big) \simeq \frac{C}{i^2} \tag{A.9}$$

for some constant $C$. This means that

$$\sum_{i=1}^{\infty}\mathrm{Var}\log(1-V_i) < \infty \tag{A.10}$$

(again using the limit comparison test). Equations (A.4) and (A.10), together with an easy application of the Chebyshev inequality, imply that $\sum_{i=1}^{n}\log(1-V_i) \xrightarrow{P} -\infty$. According to Gut, Theorem 3.3.5, convergence in probability implies convergence almost surely for a monotone sequence. The result follows.

□

**Lemma A.2.** *Let $\alpha_i > 0$ for $i \le K$ and $\sum_{i=1}^{K}\alpha_i = 1$. Let $g(p_1,\dots,p_K) = \prod_{k=1}^{K}p_k^{\alpha_k}$. Then $\sup_{\blacktriangle^K} g = \prod_{k\le K}\alpha_k^{\alpha_k}$.*

*Proof.* As $\alpha_i > 0$ for $i \leq K$, the function $g$ is continuous and, because $\blacktriangle^K$ is compact in $\mathbb{R}^K$, it achieves its extreme values. Let $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_K) \in \blacktriangle^K$ satisfy $g(\hat{\boldsymbol{p}}_K) = \sup_{\blacktriangle^K} g$. Clearly, $\hat{\boldsymbol{p}} \in \Delta^K$. Indeed, otherwise $s = \sum_{i=1}^K \hat{p}_i < 1$, $\hat{\boldsymbol{p}}/s \in \blacktriangle^K$ and $g(\hat{\boldsymbol{p}}/s) = g(\hat{\boldsymbol{p}})/s > g(\hat{\boldsymbol{p}})$, which contradicts the definition of $\hat{\boldsymbol{p}}$. Since $g$ is nonnegative on $\Delta^K$ and it is equal to 0 on the boundary of $\Delta^K$, we know that $\hat{\boldsymbol{p}}$ is in the interior of $\Delta^K$. The function $g$ is positive on the interior of $\Delta^K$, so by considering the function $\log(g)$ and using the Lagrange multipliers, we got that $\hat{\boldsymbol{p}}$ satisfies

$$0 = (\alpha_i \log p_i)' + \lambda = \frac{\alpha_i}{p_i} + \lambda \tag{A.11}$$

for $i \leq K$ and some $\lambda \in \mathbb{R}$. Hence $p_i$'s are proportional to $\alpha_i$'s, and because $\sum_{i=1}^K \alpha_i = 1$, we get that $\hat{p}_i = \alpha_i$ and the proof follows. □

**Lemma A.3.** *Let $\mathcal{Y}$ be a measurable space and $r \in \mathbb{N}$. Let $f \in L^r(\mathcal{Y}) \cap L^\infty(\mathcal{Y})$ and $f, f_n \geq 0$ for $n \geq 1$. Assume that $\|f_n\|_r \to \|f\|_r$ and $\|f_n - f\|_\infty \to 0$. Then*

$$\|f_n\|_n \to \|f\|_\infty.$$

*Proof.* Fix $\varepsilon > 0$. By Lemma 2.14 $\|f\|_n \to \|f\|_\infty$ and hence there exist $N_1$ such that

$$\big|\|f\|_n - \|f\|_\infty\big| < \varepsilon \quad \text{for } n > N_1. \tag{A.12}$$

By the assumptions there exists $N_2$ such that

$$\|f_n - f\|_\infty < \varepsilon \quad \text{for } n > N_2 \tag{A.13}$$

and, using Lemma A.4, there exist $N_3$ for which

$$\|f_n - f\|_r \leq \sqrt[r]{2}\|f\|_r + \varepsilon \quad \text{for } n > N_3. \tag{A.14}$$

It is clear that for any function $g$ and $n > r$ we have

$$\|g\|_n^n = \int_{\mathcal{Y}} |g|^n \leq \|g\|_\infty^{n-r} \int_{\mathcal{Y}} |g|^r = \|g\|_\infty^{n-r} \|g\|_r^r \tag{A.15}$$

and hence, placing $g = f_n - f$ and taking the $n$-th root

$$\|f_n - f\|_n \leq \|f_n - f\|_\infty^{\frac{n-r}{n}} \cdot \|f_n - f\|_r^{\frac{r}{n}}. \tag{A.16}$$

Pulling (A.12), (A.13), (A.14) and (A.16) together we obtain that for $n > \max\{N_1, N_2, N_3, r\}$

$$\begin{aligned} \big|\|f_n\|_n - \|f\|_\infty\big| &\leq \big|\|f_n\|_n - \|f\|_n\big| + \big|\|f\|_n - \|f\|_\infty\big| \leq \\ &\leq \|f_n - f\|_n + \varepsilon \leq \\ &\leq \varepsilon^{\frac{n-r}{n}} \big(\sqrt[r]{2}\|f\|_r + \varepsilon\big)^{\frac{r}{n}} + \varepsilon \end{aligned} \tag{A.17}$$

and hence

$$\limsup_{n\to\infty} \big|\|f_n\|_n - \|f\|_\infty\big| \leq 2\varepsilon. \tag{A.18}$$

As the choice of $\varepsilon > 0$ was arbitrary, the proof follows. □

**Lemma A.4.** *Let $\mathcal{Y}$ be a measurable space and let $r \in \mathbb{N}$. Let $f \in L^r(\mathcal{Y})$ and $f, f_n \geq 0$ for $n \geq 1$. Assume that $\|f_n\|_r \to \|f\|_r$. Then*

$$\limsup_{n\to\infty} \|f - f_n\|_r \leq \sqrt[r]{2}\|f\|_r.$$

*Proof.* We start with the case $r = 1$. For any function $g$ let $g^+(x) = g(x)\mathbf{1}_{g(x)\geq 0}$ and $g^-(x) = -g(x)\mathbf{1}_{g(x)<0}$. Let $d_n = f - f_n$. Note that

$$\|d_n\|_1 = \|d_n^+\|_1 + \|d_n^-\|_1. \tag{A.19}$$

Moreover

$$\|f_n\|_1 - \|f\|_1 = \|d_n^+\|_1 - \|d_n^-\|_1 \tag{A.20}$$

and since $f, f_n \geq 0$ we have $d_n^+ \leq f$ and

$$\|d_n^+\|_1 \leq \|f\|_1. \tag{A.21}$$

Using (A.19), (A.20) and (A.21), together with the assumption $\|f_n\|_1 \to \|f\|_1$, we get

$$\limsup_{n\to\infty} \|d_n\|_1 = \limsup_{n\to\infty} \big(2\|d_n^+\|_1 - (\|f_n\|_1 - \|f\|_1)\big) \leq 2\|f\|_1 \tag{A.22}$$

and that finishes the proof of the case $r = 1$.

For general $r \in \mathbb{N}$ we use the obvious equality $\|g\|_r^r = \|g^r\|_1$ for $g \geq 0$. From the assumptions $\|f_n^1\|_1 \to \|f^r\|_1$, which implies

$$\limsup_{n\to\infty} \|f_n^r - f^r\|_1 \leq 2\|f^r\|_1 = 2\|f\|_r^r. \tag{A.23}$$

It is clear that for any $z \geq 0$ and $r \in \mathbb{N}$ we have $z^r \leq (z+1)^r - 1$ and hence for any $x \geq y > 0$, by placing $z = \frac{x}{y} - 1$ and multiplying by $y^r$ we get $|x - y|^r \leq |x^r - y^r|$. But this inequality is symmetric with respect to $x, y$, so it is valid for any $x, y > 0$ and – by considering the obvious case $x = 0$ or $y = 0$ – for any $x, y \geq 0$. It then follows that

$$\|f_n - f\|_r^r = \int_{\mathcal{Y}} |f_n - f|^r \leq \int_{\mathcal{Y}} |f_n^r - f^r| = \|f_n^r - f^r\|_1. \tag{A.24}$$

We finish the proof by joining (A.23) and (A.24). □

**Lemma A.5.** *Let $f$ be a strictly convex function on $\mathbb{R}^k$ with values in $\mathbb{R} \cup \{-\infty, \infty\}$. Let $U$ be the essential domain of $f$, i.e. $U = \{x \in \mathbb{R}^k \colon f(x) < \infty\}$ and assume that $U$ is an open subset of $\mathbb{R}^k$. If there exists $x_0 \in U$ such that $f(x_0) = \inf_{x\in U} f(x)$ then for every $a \in \mathbb{R}$ the set $U_a := \{x \in \mathbb{R}^k \colon f(x) \leq a\}$ is a compact subset of $U$.*

*Proof.* Without loss of generality assume that $x_0 = \mathbf{0}^k$ and $f(x_0) = 0$. The theorem obviously holds for $a < 0$. Fix $a \geq 0$. Clearly $U_a \subseteq U$. As $f$ is convex, it is continuous in $U$ (cf. Rockafellar, 1970, Theorem 10.1) and hence $U_a$ is a closed set. It is left to prove that it is bounded. Take $\varepsilon > 0$ such that $B(0, \varepsilon) \subseteq U$ and let $M = \inf_{x\in\partial B(0,\varepsilon)} f(x)$. Clearly, by the convexity of $f$: $f(x) > M\frac{\|x\|}{\varepsilon}$ for $x \notin B(0, \varepsilon)$. Therefore $U_a \subseteq B(0, \frac{\varepsilon}{M})$ and the proof follows. □

## A.1 Detailed computations of marginal distributions from Section 1.4.1

### A.1.1 Normal-Normal

If $\Sigma$ is a symmetric matrix and $v = \begin{bmatrix} \mathrm{diag}(\Sigma_0) \\ \mathrm{low}(\Sigma_0) \end{bmatrix}$, let $\mathbf{M}(v) = \Sigma$. We can rewrite the formula for $\mathsf{C}(\tau, \zeta)$ in (1.33) in a more direct form as

$$\mathsf{C}(\tau, \zeta) = \frac{1}{2} \log |\mathbf{M}(\zeta)|^{-1} + \frac{1}{2} \tau^\top \mathbf{M}(\zeta)^{-1} \tau \tag{A.25}$$

We have

$$\begin{aligned} \tau_{\boldsymbol{x}} &= \Psi_0^{-1}\mu_0 + \sum_{i=1}^{k} \Sigma_0^{-1} x_i = \Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}}, \\ \zeta_k &= \begin{bmatrix} \mathrm{diag}(\Psi_0^{-1}) \\ \mathrm{low}(\Psi_0^{-1}) \end{bmatrix} + k \begin{bmatrix} \mathrm{diag}(\Sigma_0^{-1}) \\ \mathrm{low}(\Sigma_0^{-1}) \end{bmatrix} = \begin{bmatrix} \mathrm{diag}(\Psi_k^{-1}) \\ \mathrm{low}(\Psi_k^{-1}) \end{bmatrix} \end{aligned} \tag{A.26}$$

and therefore by (A.25)

$$\mathsf{C}(\tau_{\boldsymbol{x}}, \zeta_k) = \frac{1}{2} \log |\Psi_k| + \frac{1}{2} (\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}})^\top \Psi_k (\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}}). \tag{A.27}$$

Hence

$$\mathsf{C}(\tau_{\boldsymbol{x}}, \zeta_k) - \mathsf{C}(\tau, \zeta) = \frac{1}{2} \log \frac{|\Psi_k|}{|\Psi_0|} + \frac{1}{2} (\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}})^\top \Psi_k (\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}}) - \frac{1}{2} \mu_0^\top \Psi_0^{-1} \mu_0. \tag{A.28}$$

We also have

$$\prod_{i=1}^{k} h(x_i) = (2\pi)^{-dk/2} |\Sigma_0|^{-k/2} \prod_{i=1}^{k} \exp\left\{ -\frac{1}{2} x_i^\top \Sigma_0^{-1} x_i \right\} \tag{A.29}$$

Note that if $A, B$ are two invertible matrices of the same size and $C = (A+B)^{-1}$ then

$$ACA - A = ACA - AC(A+B) = -ACB = BCB - (A+B)CB = BCB - B \tag{A.30}$$

By plugging $A = \Psi_0^{-1}$ and $B = k\Sigma_0^{-1}$ into the left-hand, the middle and the right-hand parts of (A.30) we get

$$\Psi_0^{-1}\Psi_k\Psi_0^{-1} - \Psi_0^{-1} = -k\Psi_0^{-1}\Psi_k\Sigma_0^{-1} = k^2\Sigma_0^{-1}\Psi_k\Sigma_0^{-1} - k\Sigma_0^{-1} \tag{A.31}$$

Hence

$$\begin{aligned} &(\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}})^\top \Psi_k (\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}}) - \mu_0^\top \Psi_0^{-1}\mu_0 = \\ &\quad = \mu_0^\top (\Psi_0^{-1}\Psi_k\Psi_0^{-1} - \Psi_0^{-1})\mu_0 + 2k\mu_0^\top \Psi_0^{-1}\Psi_k\Sigma_0^{-1}\overline{\boldsymbol{x}} + k^2\overline{\boldsymbol{x}}\Sigma_0^{-1}\Psi_k\Sigma_0^{-1}\overline{\boldsymbol{x}} = \\ &\quad = \mu_0^\top (k^2\Sigma_0^{-1}\Psi_k\Sigma_0^{-1} - k\Sigma_0^{-1})\mu_0 - 2\mu_0^\top (k^2\Sigma_0^{-1}\Psi_k\Sigma_0^{-1} - k\Sigma_0^{-1})\overline{\boldsymbol{x}} + k^2\overline{\boldsymbol{x}}\Sigma_0^{-1}\Psi_k\Sigma_0^{-1}\overline{\boldsymbol{x}} = \\ &\quad = -k\mu_0\Sigma_0^{-1}\mu_0 + 2k\mu_0\Sigma_0^{-1}\overline{\boldsymbol{x}} + k^2\big(\overline{\boldsymbol{x}}\Sigma_0^{-1}\Psi_k\Sigma_0^{-1}\overline{\boldsymbol{x}} - 2\mu_0^\top\Sigma_0^{-1}\Psi_k\Sigma_0^{-1}\overline{\boldsymbol{x}} + \mu_0^\top\Sigma_0^{-1}\Psi_k\Sigma_0^{-1}\mu_0\big) = \\ &\quad = \Big( \sum_{i=1}^{k} x_i^\top \Sigma_0^{-1} x_i - \sum_{i=1}^{k} (x_i - \mu_0)^\top \Sigma_0^{-1} (x_i - \mu_0) \Big) + k^2 (\overline{\boldsymbol{x}} - \mu_0)^\top \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} (\overline{\boldsymbol{x}} - \mu_0) \end{aligned} \tag{A.32}$$

Plugging (A.28) and (A.29) into (1.26) gives us (1.34).

### A.1.2 Normal-Inverse-Wishart

Let $\zeta = \begin{bmatrix} \zeta^{(1)} \\ \zeta^{(2)} \end{bmatrix}$ and for $\tau = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\Lambda) \\ -\mathrm{low}(\Lambda^\top) \\ \mu \end{bmatrix}$ let $\tau^{(1)} = \begin{bmatrix} -\frac{1}{2}\mathrm{diag}(\Lambda) \\ -\mathrm{low}(\Lambda^\top) \end{bmatrix}$ and $\tau^{(2)} = \mu$. Moreover let $\tilde{\mathbf{M}}(\tau^{(2)}) = \Lambda$. Then

$$\mathsf{C}(\tau, \zeta) = -\frac{d}{2}\log \zeta^{(2)} - \frac{\zeta^{(1)} - d - 2}{2} \log \frac{\left|\tilde{\mathbf{M}}(\tau^{(1)}) - \frac{1}{\zeta^{(2)}}\left(\tau^{(2)}\right)\left(\tau^{(2)}\right)^\top\right|}{2^d} + \log \Gamma_d\left(\frac{\zeta^{(1)} - d - 2}{2}\right). \tag{A.33}$$

Note that

$$\begin{aligned}
&\tilde{\mathbf{M}}(\tau_{\boldsymbol{x}}^{(1)}) - \frac{1}{\zeta_k^{(2)}}\left(\tau_{\boldsymbol{x}}^{(2)}\right)\left(\tau_{\boldsymbol{x}}^{(2)}\right)^\top = \\
&\quad = \Big(\nu_0\Sigma_0 + \kappa_0\mu_0{\mu_0}^\top + \sum_{i=1}^d x_i{x_i}^\top\Big) - \frac{1}{\kappa_0 + k}\big(\kappa_0\mu_0 + \sum_{i=1}^d x_i\big)\big(\kappa_0\mu_0 + \sum_{i=1}^d x_i\big)^\top = \\
&\quad = \nu_0\Sigma_0 + \sum_{i=1}^d x_i{x_i}^\top - k\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^\top + k\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^\top - \frac{1}{\kappa_0 + k}\big(\kappa_0\mu_0 + k\overline{\boldsymbol{x}}\big)\big(\kappa_0\mu_0 + k\overline{\boldsymbol{x}}\big)^\top + \kappa_0\mu_0{\mu_0}^\top = \\
&\quad = \Sigma(\boldsymbol{x})
\end{aligned} \tag{A.34}$$

where $\Sigma(\boldsymbol{x})$ is given by (1.47). Putting together (A.34) and (A.33) and applying this to (1.26) yields (1.45).

### A.1.3 Normal-Inverse-Gamma

For $\tau = \begin{bmatrix} a \\ \boldsymbol{v} \end{bmatrix}$, where $a \in \mathbb{R}$, $\boldsymbol{v} \in \mathbb{R}^d$ let $\tau^{(1)} = a$, $\tau^{(2)} = \boldsymbol{v}$. Similarly, for $\zeta = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}$, where $b \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^{d(d+1)/2}$ let $\zeta^{(1)} = b$ and $\zeta^{(2)} = \boldsymbol{w}$. If $\Sigma$ is a symmetric matrix and $v = \begin{bmatrix} \mathrm{diag}(\Sigma_0) \\ \mathrm{low}(\Sigma_0) \end{bmatrix}$, let $\mathbf{M}(v) = \Sigma$. Then

$$\mathsf{C}(\tau, \zeta) = -(\zeta^{(1)} - 1)\log\Big(-\tau^{(1)} - \frac{1}{2}\tau^{(2)}\mathbf{M}(\zeta^{(2)})^{-1}\tau^{(2)\top}\Big) + \log\Gamma(\zeta^{(1)} - 1) - \frac{1}{2}\log|\mathbf{M}(\zeta^{(2)})| \tag{A.35}$$

We have

$$\mathbf{M}(\zeta_k^{(2)}) = \Psi_0^{-1} + k\Sigma_0^{-1} = \Psi_k^{-1} \tag{A.36}$$

By applying (A.31) we get

$$(\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}})^\top\Psi_k(\Psi_0^{-1}\mu_0 + k\Sigma_0^{-1}\overline{\boldsymbol{x}}) - {\mu_0}^\top\Psi_0^{-1}\mu_0 - k\overline{\boldsymbol{x}}^\top\Sigma_0^{-1}\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}} - \mu_0)^\top k\Xi_k(\overline{\boldsymbol{x}} - \mu_0) \tag{A.37}$$

and hence

$$
\begin{aligned}
&-\tau_{\boldsymbol{x}}^{(1)} - \frac{1}{2}\tau_{\boldsymbol{x}}^{(2)}\mathbf{M}(\zeta_k^{(2)})^{-1}\tau_{\boldsymbol{x}}^{(2)\top} = \\
&= \Big(\beta_0\gamma_0 + \frac{1}{2}\mu_0{}^\top\Psi_0^{-1}\mu_0 + \frac{1}{2}\sum_{i=1}^{k} x_i{}^\top\Sigma_0^{-1}x_i\Big) - \frac{1}{2}(\Psi_0^{-1}\mu_0 + k\Sigma_0\overline{\boldsymbol{x}})^\top\Psi_k(\Psi_0^{-1}\mu_0 + k\Sigma_0\overline{\boldsymbol{x}}) = \\
&= \beta(\boldsymbol{x}).
\end{aligned}
\tag{A.38}
$$

By applying (A.35), (A.36) and (A.38) to (1.26) we obtain (1.56).

## A.2 Linear algebra

**Lemma A.6.** *Symmetric, positive definite matrices have the following properties*

*(a) the sum of symmetric positive definite matrices is symmetric positive definite.*

*(b) the inverse of symmetric positive definite matrix is symmetric positive definite.*

*(c) for each symmetric positive matrix $A$ there exist an uniquely defined symmetric positive matrix $B$ such that $A = B^\top B$. We use the notation $B = A^{1/2}$.*

*(d) if $A, B$ are symmetric positive definite matrices and also $A - B$ is symmetric positive definite then $B^{-1} - A^{-1}$ is symmetric positive definite.*

*(e) if $A, B$ are positive definite then $\det(A + B) \geq \det A$.*

*Proof.* Let $A, B \in \mathbb{R}^{d,d}$.

(a) If $A, B$ are symmetric then $A + B$ is also symmetric. If $A, B$ are positive definite then for every $x \in \mathbb{R}^d \setminus \{\vec{0}\}$ we have $x^\top(A + B)x = x^\top Ax + x^\top Bx > 0$ and hence $A + B$ is also positive definite.

(b) If $A$ is symmetric then $A^{-1}$ is also symmetric. If $A$ is positive definite then by Theorem 7.1 from Zhang (2011) it may be expressed as $U^*\mathrm{diag}\lambda_1, \ldots, \lambda_d U$ where $U$ is unitary matrix and $U^*$ its conjugate transpose and $\lambda_1, \ldots, \lambda_d > 0$. Therefore $A^{-1} = U^*\mathrm{diag}\lambda_1^{-1}, \ldots, \lambda_d^{-1}U$ and again by using Theorem 7.1 we obtain that $A^{-1}$ is positive definite.

(c) Since if $A$ is a symmetric matrix then $A^\top A = A^2$ and this point is an easy consequence of Theorem 7.4 in Zhang (2011).

(d) Let $P$ be a symmetric matrix that satisfy $P^2 = B$. Positive definiteness of $A - B$ is equivalent to $x'Ax > x'Bx$ for all $x \in \mathbb{R}^d$. By substituting $y = Px$ this is equivalent to $y'P^{-1}AP^{-1}y > y'y$ for all $y \in \mathbb{R}^d$. Note that $P^{-1}AP^{-1}$ is positive definite (as a product of positive definite matrices) and hence it can be expressed as $U^*\Lambda U$. But then the latest condition can be expressed as $z'z > z'U^*\Lambda^{-1}Uz$ for all $z \in \mathbb{R}^d$ which in the same way is equivalent to the positive definiteness of $B^{-1} - A^{-1}$.

(e) This is clearly equivalent to $\det(I+BA^{-1}) \geq \det(I) = 1$. As $BA^{-1}$ is positive definite then for every eigenvalue $\lambda$ of $I + BA^{-1}$ we have $\lambda = v'(I + BA^{-1})v > 1$, where $v$ is its eigenvector of norm 1. Therefore the determinant of $I + BA^{-1}$ is also greater than 1.

□

**Lemma A.7.** *If $A$ is a square matrix then $I - (I + A^2)^{-1} = A(I + A^2)^{-1}A$.*

*Proof.* We have

$$(I + A^2)^{-1} + A^2(I + A^2)^{-1} = (I + A^2)(I + A^2)^{-1} = I, \tag{A.39}$$

and hence

$$I - (I + A^2)^{-1} = A^2(I + A^2)^{-1}. \tag{A.40}$$

Using the formula for the inverse of a product, $(AB)^{-1} = B^{-1}A^{-1}$ we get

$$\left(A^2(I + A^2)^{-1}\right)^{-1} = (I + A^2)A^{-2} = A^{-2} + I, \tag{A.41}$$

and hence $A^2(I + A^2)^{-1} = (A^{-2} + I)^{-1}$. Similarly

$$\left(A(I + A^2)^{-1}A\right)^{-1} = A^{-1}(I + A^2)A^{-1} = A^{-2} + I, \tag{A.42}$$

and the proof follows. □

**Lemma A.8.** *If $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric, positive definite then $|\Sigma| \leq \|\Sigma\|^d$, where $|.|$ denotes determinant and $\|.\|$ denotes operator norm.*

*Proof.* Let $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ be the eigenvalues of the matrix $\Sigma$. Then

$$|\Sigma| = \prod_{i=1}^{d} \lambda_i \quad \text{and} \quad \|\Sigma\| = \lambda_d, \tag{A.43}$$

and the result follows. □

**Lemma A.9.** *Let $u \in \mathbb{R}^d$. Then $\|uu^\top\| = \|u\|^2$.*

*Proof.* By the Cauchy-Schwarz inequality, for any $v \in \mathbb{R}^d$ we have

$$\|(uu^\top)v\|^2 = \left((uu^\top)v\right)^\top (uu^\top)v = (v^\top u)(u^\top u)(u^\top v) \leq \|u\|^2(u^\top v)^2 \leq \|u\|^4\|v\|^2. \tag{A.44}$$

The inequality becomes an equality for $v = u$ and the result follows. □

**Lemma A.10.** *Here we present some well-known properties of the trace. If $A, B$ are symmetric, positive definite $d \times d$ matrices then*

*(i)* $\mathrm{tr}(AB) \leq \sqrt{\mathrm{tr}(A^2)\mathrm{tr}(B^2)}$

*(ii)* $\mathrm{tr}(A^2) \leq \mathrm{tr}(A)^2$

*(iii)* $\mathrm{tr}(A^{-1}) \leq d^2 \mathrm{tr}(A)^{-1}$

*As a result:*

*(a)* $\mathrm{tr}(AB) \leq \mathrm{tr}(A)\mathrm{tr}(B)$

*(b)* $\mathrm{tr}(A^{-1}) \geq d\mathrm{tr}(A)^{-1}$.

*Proof.* Part (i) is a Cauchy-Schwarz inequality that can be applied since $\mathrm{tr}(AB)$ is a scalar product on the space of symmetric matrices, as is easily demonstrated. Part (ii) is a consequence of the fact that the trace is a sum of eigenvalues, which are positive for positive definite matrices, and that the eigenvalues of $A^2$ are the squares of eigenevalues of $A$. Part (iii) is the inequality between arithmetic and harmonic means applied to the eigenvalues of $A$. Part (a) follows easily from parts (i) and (ii), and part (b) follows from (a) by setting $B = A^{-1}$. □

## A.3 Proof of Lemma 3.15

This is a straightforward consequence of the definition $E_n(X \mid X \in C) = E_n X \mathbf{1}_{X \in C} / P_n(C)$ and the following Lemma A.11.

**Lemma A.11.** *If $P$ satisfies $(*)$ and for $X \sim P$ we have $E_P \|X\|^2 < \infty$ then $P$ satisfies*

$$\lim_{n\to\infty} \sup_{C \text{ convex}} \left\| E_n X \mathbf{1}_{X \in C} - E_P X \mathbf{1}_{X \in C} \right\| = 0 \quad \textit{almost surely.} \tag{$**$}$$

*where $E_n f(X) = \int_{\mathcal{X}} f(X) \mathrm{d}P_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$.*

*Proof.* Let $\mathcal{K}$ be the space of all convex sets. If $A$ is a set, then $\mathcal{K} \cap A := \{K \cap A \colon K \in \mathcal{K}\}$. Let $x^{(i)}$ $(i \leq d)$ be the $i$-th coordinate of vector $x$. We now prove that for every $r > 0$

$$\lim_{n\to\infty} \sup_{C \in \mathcal{K} \cap [-r,r]^d} \left| E_n X^{(1)} \mathbf{1}_{X \in C} - E_P X^{(1)} \mathbf{1}_{X \in C} \right| = 0. \tag{A.45}$$

Fix $r > 0$ and $C \in \mathcal{K} \cap [-r, r]^d$. For $m \in \mathbb{N}$ and $-m \leq k \leq m-1$ let $C_k^m = C \cap [rk/m, r(k+1)/m) \times \mathbb{R}^{d-1}$. Then

$$\left| E_P X^{(1)} \mathbf{1}_{X \in C} - \sum_{k=-m}^{m-1} r \frac{k}{m} P(C_k^m) \right| \leq \frac{r}{m} P(C) \leq \frac{r}{m}. \tag{A.46}$$

It follows from the same reasoning

$$\left| E_n X^{(1)} \mathbf{1}_{X \in C} - \sum_{k=-m}^{m-1} r \frac{k}{m} P_n(C_k^m) \right| \leq \frac{r}{m} \quad \text{for every } n \in \mathbb{N}. \tag{A.47}$$

Now choose $\varepsilon > 0$ and $m > r/\varepsilon$. Note that $C_k^m$ are convex sets (as intersections of convex sets) and hence by $(*)$ we may choose $N$ so that for $n > N$ and any convex $C'$ we have that $|P_n(C') - P(C')| < \varepsilon/(2m)$ and hence

$$\begin{aligned}\Big|\sum_{k=-m}^{m-1} r\frac{k}{m}P(C_k^m) - \sum_{k=-m}^{m-1} r\frac{k}{m}P_n(C_k^m)\Big| &\leq \sum_{k=-m}^{m-1} r\frac{|k|}{m}|P(C_k^m) - P_n(C_k^m)| < \\ &< \sum_{k=-m}^{m-1} r\frac{\varepsilon}{2m} < r\varepsilon.\end{aligned} \tag{A.48}$$

By combining (A.46), (A.47) and (A.48) we obtain that $|E_n X^{(1)}\mathbf{1}_{X\in C} - E_P X^{(1)}\mathbf{1}_{X\in C}| < (2+r)\varepsilon$ for $n > N$ and since the choice of $N$ does not depend on $C$, (A.45) follows.

We now prove that almost surely

$$\lim_{n\to\infty}\sup_{C\in\mathcal{K}}\big|E_n X^{(1)}\mathbf{1}_{X\in C} - E_P X^{(1)}\mathbf{1}_{X\in C}\big| = 0. \tag{A.49}$$

The same result for the remaining coordinates of $(**)$ follows in the same way, from which follows the statement of the Lemma. Note that the function $r \mapsto E_P|X^{(1)}|\mathbf{1}_{X\notin[-r,r]^d}$ is decreasing to 0 as $r$ goes to infinity. By the Strong Law of Large Numbers almost surely $\lim_{n\to\infty} E_n|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} = E_P|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d}$ for every $K \in \mathbb{N}$.

Fix $C \in \mathcal{K}$ and $\varepsilon > 0$. Since $\lim_{K\to\infty} E_P|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} = 0$ it follows that there exist $K \in \mathbb{N}$ such that $E_P|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} < \varepsilon$ and $\lim_{n\to\infty} E_n|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} < \varepsilon$. The latter means that there exist $n_1$ such that $E_n|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} < \varepsilon$ for every $n > n_1$. By (A.45) there exist $n_2 \in \mathbb{N}$ such that for every $n > n_2$

$$\big|E_n X^{(1)}\mathbf{1}_{X\in C\cap[-K,K]^d} - E_P X^{(1)}\mathbf{1}_{X\in C\cap[-K,K]^d}\big| < \varepsilon. \tag{A.50}$$

Therefore for $n > \max\{n_1, n_2\}$ we get

$$\begin{aligned}\big|E_n X^{(1)}\mathbf{1}_{X\in C} - E_P X^{(1)}\mathbf{1}_{X\in C}\big| < \big|E_n X^{(1)}\mathbf{1}_{X\in C\cap[-K,K]^d} - E_P X^{(1)}\mathbf{1}_{X\in C\cap[-K,K]^d}\big| + \\ + E_n|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} + E_P|X^{(1)}|\mathbf{1}_{X\notin[-K,K]^d} < 3\varepsilon\end{aligned} \tag{A.51}$$

Because $n_1, n_2$ do not depend on $C$, (A.49) follows, which finishes the proof of the Lemma. □

**Lemma A.12.** *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.* Assume that $(\mathcal{F}, d)$ is a (pseudo)metric space. We prove that $(F_K(\mathcal{F}), \bar{d})$ is also a (pseudo)metric space. Take any $\mathcal{A} = \{A^{(1)}, \ldots, A^{(k)}\} \in F_K(\mathcal{F})$ and $\mathcal{B} = \{B^{(1)}, \ldots, B^{(l)}\} \in F_K(\mathcal{F})$. By definition

$$\bar{d}(\mathcal{A}, \mathcal{B}) = \min_{\sigma\in\Sigma_K}\max_{i\leq K} d(A^{(i)}, B^{(\sigma(i))}) \geq 0, \tag{A.52}$$

since $d(A^{(i)}, B^{(j)}) \geq 0$ for any $i, j \leq K$ (as in the definition we assume that $A^{(i)} = \emptyset$ and $B^{(j)} = \emptyset$ for $i > k$ or $j > l$ respectively). Let $\mathcal{C} = \{C^{(1)}, \dots, C^{(l)}\} \in F_K(\mathcal{F})$ and let $\sigma_1$, $\sigma_2$ and $\sigma_3$ be permutations of $[K]$ that satisfy

$$\overline{d}(\mathcal{A}, \mathcal{B}) = \max_{i \leq K} d(A^{(i)}, B^{(\sigma_1(i))}) \quad \text{and} \quad \overline{d}(\mathcal{B}, \mathcal{C}) = \max_{i \leq K} d(B^{(i)}, C^{(\sigma_2(i))}) \tag{A.53}$$

Note that $d(A^{(i)}, B^{(\sigma_1(i))}) + d(B^{(\sigma_1(i))}, C^{(\sigma_2(\sigma_1(i)))}) \geq d(A^{(i)}, C^{(\sigma_2(\sigma_1(i)))})$ and hence

$$\begin{aligned} \overline{d}(\mathcal{A}, \mathcal{B}) + \overline{d}(\mathcal{B}, \mathcal{C}) &= \max_{i \leq K} d(A^{(i)}, B^{(\sigma_1(i))}) + \max_{i \leq K} d(B^{(\sigma_1(i))}, C^{(\sigma_2(\sigma_1(i)))}) \geq \\ &\geq \max_{i \leq K} \left( d(A^{(i)}, B^{(\sigma_1(i))}) + d(B^{(\sigma_1(i))}, C^{(\sigma_2(\sigma_1(i)))}) \right) \geq \\ &\geq \max_{i \leq K} d(A^{(i)}, C^{(\sigma_2 \circ \sigma_1(i))}) \geq \overline{d}(\mathcal{A}, \mathcal{C}) \end{aligned} \tag{A.54}$$

and the triangle inequality follows. This means that $\overline{d}$ is a pseudometric on $\mathcal{F}_K$.

Now assume that $(\mathcal{F}, d)$ is finitely compact. Let $(\mathcal{A}_n)_{n=1}^{\infty}$ be a sequence in $F_K(\mathcal{F})$ and let $\mathcal{A}_n = \{A_n^{(1)}, A_n^{(2)}, \dots, A_n^{(k_n)}\}$. As the sequence $(k_n)_{n=1}^{\infty}$ is bounded by $K$ we may choose a subsequence $\mathcal{A}_{n_k}$ and $\tilde{K} \in \mathbb{N}$ such that $|\mathcal{A}_{n_k}| = \tilde{K}$ for every $k \in \mathbb{N}$. Consider the sequence $(A_{n_k}^{(1)})_{k=1}^{\infty}$. This sequence is bounded (as $(\mathcal{A}_n)_{n=1}^{\infty}$ is bounded). Therefore it has a subsequence $(A_{n_{k_l}}^{(1)})_{l=1}^{\infty}$ converging in $d$ to $A^{(1)} \in \mathcal{F}$. Now we consider $(A_{n_{k_l}}^{(2)})_{l=1}^{\infty}$ and again we choose a subsequence $(A_{n_{k_{l_m}}})_{m=1}^{\infty}$ converging in $d$ to $A^{(2)} \in \mathcal{F}$. By iterating this procedure $\tilde{K}$ times we obtain a family $\hat{A} = \{A^{(1)}, \dots, A^{(\tilde{K})}\}$ of 'limiting' sets. It is easy to verify that the final subsequence of $(\mathcal{A}_n)_{n=1}^{\infty}$ converges in $\overline{d}$ to $\hat{A}$, which finishes the proof.

□

**Lemma A.13.** *If $P$ is a measure on $(\mathbb{R}^d, \mathcal{B})$ with bounded support and absolutely continuous with respect to the Lebesgue measure then $\Psi(\alpha) > 0$ for every $\alpha > 0$, where $\Psi$ is given by (3.79).*

*Proof.* Fix $\alpha > 0$. As an easy consequence of Theorem 3.19 we obtain that $P(\cdot)$ is a continuous function in $(\mathcal{K}_r, \varrho_H)$. Therefore $\mathcal{K}_r^{\alpha} := \{A \in \mathcal{K}_r \colon P(A) \geq \alpha\}$ is a closed (as a counterimage of closed half-line in continuous transformation) subspace of compact (by Theorem 3.18) topological space, therefore it is compact itself.

Assume that the support of $P$ is contained in the ball $B(\mathbf{0}, r)$ and without losing generality assume that $r > 1$. Consider the function

$$\varphi(A) = \sup_{\substack{A_1, A_2 \in \mathcal{B} \\ A_1 \cup A_2 = A \\ A_1 \cap A_2 = \emptyset}} P(A_1) \cdot P(A_2) \cdot \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\|^2 \tag{A.55}$$

in the compact topological space $(\mathcal{K}_r, \varrho_H)$. We prove that this function is continuous.

Firstly note that if $C \in \mathcal{K}_r^{\alpha}$ then $\|E_P X \mathbf{1}_{X \in C}\| \leq r P(C)$. From this it can be easily seen that for every $\varepsilon > 0$ there exist $\delta > 0$ such that for $A, B \in \mathcal{K}_r^{\alpha}$ if $d_P(A, B) < \delta$ then $\|E_P(X \,|\, X \in A) - E_P(X \,|\, X \in B)\| < \varepsilon$.

Fix $0 < \varepsilon < 1$. There exist $\delta_1 < \varepsilon$ such that if $A, A' \in \mathcal{K}_r^\alpha$ and $d_P(A, A') < \delta_1$ then $\|E_P(X \,|\, X \in A) - E_P(X \,|\, X \in A')\| < \varepsilon/2$. There exist $\delta_2$ such that if $A, A' \in \mathcal{K}_r$ and $\varrho_H(A, A') < \delta_2$ then $d_P(A, A') < \delta_1$ (this is because of Theorem 3.19 and the fact that $(\mathcal{K}_r^\alpha, \varrho_H)$ is compact and therefore the continuity implies the uniform continuity). Let us take $A, A' \in \mathcal{K}_r^\alpha$ such that $\varrho_H(A, A') < \delta_2$. Let $A_1, A_2 \in \mathcal{K}_r^\alpha$ be such that $A_1 \cap A_2 = \emptyset$, $A_1 \cup A_2 = A$ and

$$\varphi(A) - \varepsilon \leq P(A_1) \cdot P(A_2) \cdot \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\|^2 \tag{A.56}$$

Consider $A_1' = (A' \setminus A_2) \cup A_1$ and $A_2' = A' \cap A_2$. Then $A_1' \cap A_2' = \emptyset$, $A_1' \cup A_2' = A'$ and

$$d_P(A_1, A_1'), d_P(A_2, A_2') \leq d_P(A, A') \leq \delta_1. \tag{A.57}$$

Therefore $|P(A_i) - P(A_i')| < \delta_1 < \varepsilon$, $\|E_P(X \,|\, X \in A_i) - E_P(X \,|\, X \in A_i')\| < \varepsilon/2$ for $i = 1, 2$. This implies that

$$\begin{aligned}
&\|E_P(X \,|\, X \in A_1') - E_P(X \,|\, X \in A_2')\| \leq \|E_P(X \,|\, X \in A_1') - E_P(X \,|\, X \in A_1)\| + \\
&\quad + \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\| + \|E_P(X \,|\, X \in A_2) - E_P(X \,|\, X \in A_2')\| \leq \\
&\quad \leq \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\| + \varepsilon.
\end{aligned} \tag{A.58}$$

Since $p_i := |P(A_i)| \leq 1$ and $d := \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\| \leq 2r$, we get

$$\begin{aligned}
&\big|P(A_1') \cdot P(A_2') \cdot \|E_P(X \,|\, X \in A_1') - E_P(X \,|\, X \in A_2')\|^2 - \\
&\qquad - P(A_1) \cdot P(A_2) \cdot \|E_P(X \,|\, X \in A_1) - E_P(X \,|\, X \in A_2)\|^2\big| < \\
&\quad < (p_1 + \varepsilon)(p_2 + \varepsilon)(d + \varepsilon)^2 - p_1 p_2 d = \\
&\quad = d^2(p_1\varepsilon + p_2\varepsilon + \varepsilon^2) + (2d\varepsilon + \varepsilon^2)(p_1 + \varepsilon)(p_2 + \varepsilon) < 32r^2\varepsilon.
\end{aligned} \tag{A.59}$$

By (A.56) and (A.59) we obtain

$$\varphi(A) - \varepsilon - 32r^2\varepsilon \leq P(A_1') \cdot P(A_2') \cdot \|E_P(X \,|\, X \in A_1') - E_P(X \,|\, X \in A_2')\|^2 \leq \varphi(A'). \tag{A.60}$$

By symmetry we get $\varphi(A') - \varepsilon - 32r^2\varepsilon \leq \varphi(A)$ which means that $|\varphi(A) - \varphi(A')| < (1 + 32r^2)\varepsilon$ for $\varrho_H(A, A') < \delta_2$ which proofs the continuity of $\varphi$ in the topological space $(\mathcal{K}_r^\alpha, \varrho_H)$. Therefore by Weierstrass Theorem we get that

$$\inf_{\substack{A \in \mathcal{K}_r \\ P(A) \geq \alpha}} \varphi(A) = \varphi(A_0) \tag{A.61}$$

for some $A_0 \in \mathcal{K}_r$ such that $P(A_0) \geq \alpha$. It is easy to see that $\varphi(A_0) > 0$ (it is enough to divide $A_0$ into two subsets of positive measure by a hyperplane so that the center of masses of two parts do not coincide) and the Lemma follows. □

**Lemma A.14.** *Let $A \cap B = \emptyset, C := A \cup B$. Then*

$$P(A)\mathbf{V}_P(A) + P(B)\mathbf{V}_P(B) \preceq P(C)\mathbf{V}_P(C) \tag{A.62}$$

*where $\preceq$ is the Löwner partial order, i.e. $M_1 \preceq M_2$ iff $M_2 - M_1$ is non-negative definite.*

*Proof.* Let $e_1(A) = \mathbb{E}\, X \mathbf{1}_A(X)$ and $e_2(A) = \mathbb{E}\, XX^\top \mathbf{1}_A(X)$ where $X \sim P$. Then

$$\mathbf{V}_P(A) = \frac{e_2(A)}{P(A)} - \frac{e_1(A)e_1(A)^\top}{P(A)^2}. \tag{A.63}$$

Note that the functions $P, e_1, e_2$ are additive, hence

$$\begin{aligned}
&P(C)\mathbf{V}_P(C) - P(A)\mathbf{V}_P(A) - P(B)\mathbf{V}_P(B) = \\
&\quad = \left(e_2(C) - \frac{e_1(C)e_1(C)^\top}{P(C)}\right) - \left(e_2(A) - \frac{e_1(A)e_1(A)^\top}{P(A)}\right) - \left(e_2(B) - \frac{e_1(B)e_1(B)^\top}{P(B)}\right) = \\
&\quad = \frac{e_1(A)e_1(A)^\top}{P(A)} + \frac{e_1(B)e_1(B)^\top}{P(B)} - \frac{e_1(C)e_1(C)^\top}{P(C)} = \\
&\quad = \frac{e_1(A)e_1(A)^\top}{P(A)} + \frac{e_1(B)e_1(B)^\top}{P(B)} - \frac{\big(e_1(A)+e_1(B)\big)\big(e_1(A)+e_1(B)\big)^\top}{P(A)+P(B)} = \\
&\quad = \frac{P(A)P(B)}{(P(A)+P(B))}\left(\frac{e(A)}{P(B)} - \frac{e(B)}{P(A)}\right)\left(\frac{e(A)}{P(B)} - \frac{e(B)}{P(A)}\right)^\top.
\end{aligned} \tag{A.64}$$

The last matrix in (A.64) is clearly non-negative definite and the proof follows. □

But firstly we introduce a lemma which collects some key equalities and asymptotic results that will be used heavily in the proofs.

**Lemma A.15.** *For any $x > d$ and $z > 0$*

$$\left(\frac{x+z-d}{2e}\right)^{\frac{zd}{2}} < \frac{\Gamma_d\left(\frac{x+z}{2}\right)}{\Gamma_d\left(\frac{x}{2}\right)} < \left(\frac{x+z}{2}\right)^{\frac{zd}{2}}. \tag{A.65}$$

*Proof.* As proved in Kečkić and Vasić (1971) for $y > x > 1$

$$\frac{y^{y-1}}{x^{x-1}}e^{x-y} < \frac{\Gamma(y)}{\Gamma(x)} < \frac{y^{y-\frac{1}{2}}}{x^{x-\frac{1}{2}}}e^{x-y}. \tag{A.66}$$

Hence for any $u = x-d+1, x-d+2, \ldots, x$

$$\frac{\Gamma\left(\frac{u+z}{2}\right)}{\Gamma\left(\frac{u}{2}\right)} > \frac{\left(\frac{u+z}{2}\right)^{\frac{u+z}{2}-1}}{\left(\frac{u}{2}\right)^{\frac{u}{2}-1}}e^{-\frac{z}{2}} = \left(1+\frac{z}{u}\right)^{\frac{u}{2}-1}\left(\frac{u+z}{2}\right)^{\frac{z}{2}}e^{-\frac{z}{2}} > \left(\frac{u+z}{2e}\right)^{\frac{z}{2}} > \left(\frac{x+z-d}{2e}\right)^{\frac{z}{2}}. \tag{A.67}$$

and

$$\begin{aligned}
\frac{\Gamma\left(\frac{u+z}{2}\right)}{\Gamma\left(\frac{u}{2}\right)} &< \frac{\left(\frac{u+z}{2}\right)^{\frac{u+z}{2}-\frac{1}{2}}}{\left(\frac{u}{2}\right)^{\frac{u}{2}-\frac{1}{2}}}e^{-\frac{z}{2}} = \left(1+\frac{z}{u}\right)^{\frac{u}{2}}\left(1+\frac{z}{u}\right)^{-\frac{1}{2}}\left(\frac{u+z}{2}\right)^{\frac{z}{2}}e^{-\frac{z}{2}} < \\
&< e^{\frac{z}{2}}\left(1+\frac{z}{x}\right)^{-\frac{1}{2}}\left(\frac{x+z}{2}\right)^{\frac{z}{2}}e^{-\frac{z}{2}} = \left(\frac{x+z}{x}\right)^{-\frac{1}{2}}\left(\frac{x+z}{2}\right)^{\frac{z}{2}} < \left(\frac{x+z}{2}\right)^{\frac{z}{2}}.
\end{aligned} \tag{A.68}$$

The proof follows from (A.67), (A.68) and the definition of the multivariate Gamma function. □

# Bibliography


Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government Printing Office, 1970.

David J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

Horst Alzer. On some inequalities for the Gamma and Psi functions. *Mathematics of Computation*, 66(217):373–389, 1997.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

Thomas Bayes. LII. An essay towards solving a problem in the „Doctrine of Chances". By the late Rev. Mr. Bayes, FrS communicated by Mr. Price, in a letter to John Canton, AmFr S. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

David Blackwell, James B. MacQueen, et al. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Georges Darmois. Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85, 1935.

Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26, 1986a.

Persi Diaconis and David Freedman. On inconsistent Bayes estimates of location. *The Annals of Statistics*, pages 68–87, 1986b.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, pages 269–281, 1979.

Kjell Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, pages 183–201, 1974.

Joseph L. Doob. *Measure Theory.* Graduate Texts in Mathematics 143. Springer-Verlag New York, 1 edition, 1994.

J Elker, David Pollard, and Winfried Stute. Glivenko-Cantelli theorems for classes of convex sets. *Advances in Applied Probability*, 11(4):820–833, 1979.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 3 edition, 1968.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Stephen E. Fienberg et al. When did Bayesian inference become „Bayesian"? *Bayesian Analysis*, 1(1):1–40, 2006.

Andrew Gelman, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2013.

N.R. Goodman. The distribution of the determinant of a complex Wishart distributed matrix. *The Annals of Mathematical Statistics*, 34(1):178–180, 1963.

Bai-Ni Guo and Feng Qi. Refinements of lower bounds for polygamma functions. *Proceedings of the American Mathematical Society*, pages 1007–1015, 2013.

Allan Gut. *Probability: a graduate course*, volume 75. Springer Science & Business Media, 2013.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

Jovan D. Kečkić and Petar M. Vasić. Some inequalities for the Gamma function. *Publications de l'Institut Mathématique*, 11(31):107–114, 1971.

Morris Kline. *Mathematical Thought From Ancient to Modern Times*, volume 1. OUP USA, 1990.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, 1936.

Caroline Lawless and Julyan Arbel. A simple proof of Pitman–Yor's Chinese Restaurant Process from its stick-breaking representation. *Dependence Modeling*, 7(1):45–52, 2019.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

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.

Maria Moszyńska. *Selected Topics in Convex Geometry*. Birkhäuser Boston, 1 edition, 2005.

Kevin P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1(2$\sigma$2): 16, 2007.

Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 567–579. Cambridge University Press, 1936.

Jim Pitman. Combinatorial Stochastic Processes. Technical report, Lecture notes for St. Flour course, 2002.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

S. James Press. *Applied multivariate analysis: using Bayesian and frequentist methods of inference*. Courier Corporation, 2005.

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

R. Tyrrell Rockafellar. *Convex Analysis*. Number 28. Princeton university press, 1970.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1 (804):801, 1956.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Frederick A. Valentine. *Convex Sets*. McGraw-Hill Book Company, 1964.

Sara Wade, Zoubin Ghahramani, et al. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

Fuzhen Zhang. *Matrix Theory: Basic Results and Techniques*. Universitext. Springer-Verlag New York, 2 edition, 2011.