A note on the geometry of the MAP partition in some Normal Bayesian Mixture Models

Lukasz Rajkowski
John Noble
February 5, 2019

Abstract

We investigate the geometry of the maximal a posteriori (MAP) partition in the Bayesian Mixture Model where the component distribution is multivariate Normal with Normal-inverse-Wishart prior on the component mean and covariance. We prove that in this case the clusters in any MAP partition are quadratically separable. Basically this means that every two clusters are separated by a quadratic surface. In connection with results of Rajkowski [2018], where the linear separability of clusters in the Bayesian Mixture Model with a fixed component covariance matrix was proved, it gives a nice Bayesian analogue of the geometric properties of Fisher Discriminant Analysis (LDA and QDA). We also describe a simple model where the covariance shape is fixed but there is a scaling parameter which may change from cluster to cluster. We prove that in any MAP partition for this model every two clusters are separated by an ellipsoid.

1 Introduction

In the standard setting of Bayesian Mixture Models we assume that the target distribution is a random mixture of distributions from some parametrized family. We assume that the probabilities of components are sampled from a (perhaps infinitely dimensional) simplex and the parameters of the component distribution are sampled independently for each component (Tatarinova and Schumitzky [2015]). A prominent example is the Dirichlet Process Mixture Model (Antoniak [1974]), where the prior distribution on the probability weights is Sethuraman’s stick breaking process (Sethuraman [1994]).

A popular choice of the component distribution is multivariate Normal, which gives Normal Bayesian Mixture Model. There are two standard conjugate prior distributions on the component mean and covariance matrix (Gelman et al. [2013], Chapter 3.6): Normal distribution on the mean with fixed component covariance matrix or Normal-inverse-Wishart distribution (here the terminology is adopted from Murphy [2007]), where the component covariance matrix follows the inverse-Wishart distribution and the component mean (conditioned on the component covariance matrix) is Normal. The exact specification of these priors is given in Section 2.2.

Bayesian Mixture Models give a basis for cluster analysis. Indeed, one can translate the distribution on component probabilities into a discrete prior distribution on the possible data partitions – just like Sethuraman’s construction translates in to the Generalised Pólya Urn Scheme (Blackwell et al. [1973]), also known as the Chinese Restaurant Process (Aldous [1985]). The inference about clusters is based on the posterior distribution on the space of partitions. Analysing partition which maximises the posterior probability (the MAP partition) seems to be a natural choice.
In Rajkowski [2018] it is proved that in the Normal Bayesian Mixture Model, when the component covariance matrix is fixed and the prior on the component mean is Normal, the MAP partition is *convex*, i.e. the convex hulls of clusters are disjoint. An equivalent formulation is: for every two clusters in the MAP there exists a hyperplane that separates them.

Placing an inverse-Wishart prior on the cluster covariance structure, with covariances for different clusters, independent of each other, gives better modelling possibilities, since it is unusual for the covariance to be known a priori and the same for different clusters. It would be of interest to characterise the boundaries of the MAP partition in this case. Since cluster covariance structures are no longer fixed, we might expect quadratic boundaries, analogously to the Fisher’s Quadratic Discriminant Analysis (Friedman et al. [2001]). This is indeed what happens and it is the goal of the article to prove this.

2 The Setup

2.1 Bayesian Mixture Models and the MAP partition

Let $\Theta \subset \mathbb{R}^p$ be the parameter space and \{\(g_\theta: \theta \in \Theta\}\} be a family of probability densities on the observation space $\mathbb{R}^d$. Consider a prior distribution on $\Theta$ given by its density $\pi$. Let $P$ be a probability distribution on the $m$-dimensional simplex $\Delta^m = \{p = (p_i)_{i=1}^m: \sum_{i=1}^m p_i = 1$ and $p_i \geq 0$ for $i \leq m\}$ (where $m \in \mathbb{N} \cup \{\infty\}$).

$$p = (p_i)_{i=1}^m \sim \mathcal{P}$$
$$\theta = (\theta_i)_{i=1}^m \overset{iid}{\sim} \pi$$
$$x = (x_1, \ldots, x_n) \mid p, \theta \overset{iid}{\sim} \sum_{i=1}^m p_i g_{\theta_i}.$$  \hfill (2.1)

This is a *Bayesian Mixture Model*. It can model possible clusters within data; they are defined by deciding which $g_{\theta_i}$ generated a given data point. In order to formally define the clusters, we need to rewrite (2.1) as

$$p = (p_i)_{i=1}^m \sim \mathcal{P}$$
$$\theta = (\theta_i)_{i=1}^m \overset{iid}{\sim} \pi$$
$$\phi = (\phi_1, \ldots, \phi_n) \mid p, \theta \overset{iid}{\sim} \sum_{i=1}^m p_i g_{\theta_i}$$
$$x_i \mid p, \theta, \phi \sim g_{\theta_i},$$  \hfill (2.2)

Then the clusters are the classes of abstraction of the equivalence relation $i \sim j \equiv \phi_i = \phi_j$.

In this way the distribution on $m$ dimensional simplex generates a probability distribution on the partitions of set $[n]$ into at most $m$ subsets. According to Pitman [2002] this leads to an *exchangeable partition*, i.e. a random partition whose probability function is invariant with respect to permutations of indices.

**Definition 2.1.** We say that $\Pi$ is an exchangeable random partition of $[n]$ if for every partition $\mathcal{I}$ of $[n]$ and permutation $\sigma: [n] \to [n]$.

$$P(\Pi = \mathcal{I}) = P\left(\Pi = \{\{\sigma(i): i \in I\}: I \in \mathcal{I}\}\right).$$  \hfill (2.3)

In order to indicate that $\Pi$ is an exchangeable random partition of $[n]$ we use a generic notation $\Pi \sim \text{ERP}_n$. Moreover we use the notation $p_n(\mathcal{I}) := P(\Pi = \mathcal{I})$.

For $\theta \sim \pi, k \in \mathbb{N}$ and $u = (u_1, \ldots, u_k) \mid \theta \overset{iid}{\sim} g_\theta$ let $f_k$ be the resulting marginal distribution on $u$, i.e.

$$f_k(u_1, \ldots, u_k) := \int_{\Theta} \pi(\theta) \prod_{i=1}^k f_\theta(u_i) d\theta.$$  \hfill (2.4)
Let \( ERP_n \) be the exchangeable probability distribution on the space of partitions generated by \( P \). We see that (2.1) is equivalent to

\[
\mathcal{I} \sim ERP_n \quad \text{independently for all } I \in \mathcal{I}.
\]

We stress the fact that the independent sampling on the ‘lower’ level of (2.5) relates to the independence between clusters (conditioned on the random partition); within one cluster the observations are (marginally) dependent. To make the notation more concise we define

\[
f(x|I) := \prod_{I \in I} f_{|I|}(x_I).
\]

Then (2.5) becomes

\[
\mathcal{I} \sim ERP_n \quad x|I \sim f(\cdot|I).
\]

Example 2.2. Consider the Dirichlet Process Mixture Model (Antoniak [1974]). Let \( \alpha > 0 \), \( G_0 \) be a probability measure on \( \Theta \) with density \( \pi \) and \( DP(\alpha,G_0) \) be the Dirichlet Process on \( \Theta \) (Ferguson [1973]). The Dirichlet Process Mixture Model is defined by

\[
G \sim DP(\alpha,G_0) \quad x_i|G,\phi \sim g_{\phi_i} \quad \text{independently for all } i \leq n.
\]

Let \( V_1, V_2, \ldots \sim \text{Beta}(1,\alpha), p_1 = V_1, p_k = V_k \prod_{i=1}^{k-1}(1 - V_i) \) for \( k > 1 \). By Sethuraman [1994] by setting \( P \) to be the distribution of \( p \) we get that (2.8) is equivalent to (2.1). The exchangeable random partition that \( P \) generates is the Generalized Polya Urn Scheme (Blackwell et al. [1973]) or the Chinese Restaurant Process (Aldous [1985]) with the probability weight given by

\[
p_n(I) = \frac{\alpha^{|I|} \prod_{I \in I}(|I| - 1)!}{\alpha^{n} \prod_{I \in I}(|I| - 1)!},
\]

where \( \alpha^{(n)} = \alpha(\alpha + 1) \ldots (\alpha + n - 1) \). Again, setting \( ERP_n \) to be the Chinese Restaurant Process with parameter \( \alpha \) we get the equivalence between (2.8) and (2.7).

Definition 2.3. Let \( x = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n \). We say that a partition \( \hat{I} \) of \([n]\) is a MAP of \( x \) if for any other partition \( I \) of \([n]\):

\[
p_n(I)f(x|I) \leq p_n(\hat{I})f(x|\hat{I}).
\]

Notation. Here and below, \( \arg\max_{a \in A} \phi(a) \) is the set of maximisers of function \( \phi \) on the set \( A \) (note that the maximiser may not be unique). Hence the MAP partition of \( x \) in a Bayesian Mixture Model (2.7) can be defined by

\[
\hat{I} \in \arg\max_{\text{partitions } I \text{ of } [n]} p_n(I)f(x|I).
\]

2.2 Specification of the Normal Bayesian Mixture Models

In the paper we consider Normal Bayesian Mixture Models in which the component distributions are multivariate Normal. This means that the parameter space \( \Theta \) is the product of \( \mathbb{R}^d \) (component mean \( \mu \)) and the space of symmetric, positive definite, \( d \times d \) matrices \( S^+ \) (component covariance matrix \( \Lambda \)). We investigate the following three conjugate choices of prior distribution on \( \Theta \): Normal-inverse-Wishart, Normal (fixed covariance) and Normal-inverse-Gamma. In each prior we compute the expected within-group covariance matrix (which is \( E\Lambda \)) and the between group covariance matrix (i.e. \( V(\mu) \)). Moreover we present the marginal density of a random vector \( u = (u_1, \ldots, u_k) \) | \( (\mu, \Lambda) \sim \text{N}(\mu, \Lambda) \).
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. However, the meaning of this symbol is always clear from the context.

2.2.1 Normal-inverse-Wishart

\[
\Lambda \sim W^{-1}(\eta_0 + d + 1, \eta_0 \Sigma_0)
\]
\[
\mu \mid \Lambda \sim N(\mu_0, \Lambda/\kappa_0)
\]

(2.12)

Here the hyperparameters are $\kappa_0, \eta_0 > 0$, $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0 \in S^+$. This prior is listed in Gelman et al. [2013] with a slightly different hyperparameters, but we made this modification to obtain

\[
E \Lambda = \Sigma_0,
\]
\[
V(\mu) = E V(\mu \mid \Lambda) + V(\mu) = E \Lambda/\kappa_0 + V(\mu_0) = \Sigma_0/\kappa_0.
\]

(2.13)

which is consistent with the remaining two priors.

**Proposition 2.4.** Let $\theta = (\mu, \Lambda)$ have the distribution given by (2.12) and let $u = (u_1, \ldots, u_k) \mid \theta \sim N(\mu, \Lambda)$. Then the marginal distribution of $u$ is given by

\[
f_{k \mid N \mid W}(u) = \frac{|\eta_0 \Sigma_0|^{\nu_0/2} \kappa_0^{1/2} \Gamma_d(\frac{k}{2})}{\pi^{kd/2} \kappa_0^{1/2} \Gamma_d(\frac{k}{2})} \cdot \det(\Sigma(u))^{-\nu_k/2},
\]

(2.14)

where $\Gamma_d$ is the multivariate Gamma function and

\[
\nu_k = \eta_0 + d + 1 + k, \kappa_k = \kappa_0 + k \quad \text{and}
\]
\[
\Sigma(u) = \eta_0 \Sigma_0 + \sum_{i=1}^{k} (u_i - \overline{u})(u_i - \overline{u})^t + \frac{\kappa_0 k}{\kappa_k} (\overline{u} - \mu_0)(\overline{u} - \mu_0)^t.
\]

(2.15)

(2.16)

**Proof.** The proof follows from Murphy [2007], equation (266). \qed

2.2.2 Normal (fixed covariance)

\[
\Lambda = \Sigma_0
\]
\[
\mu \sim N(\mu_0, \Psi_0)
\]

(2.17)

Here the hyperparameters are $\mu_0 \in \mathbb{R}^d$ and $\Psi_0, \Sigma_0 \in S^+$. This prior is also listed in Gelman et al. [2013]. Clearly

\[
E \Lambda = \Sigma_0, \quad V(\mu) = \Psi_0.
\]

(2.18)

**Proposition 2.5.** Let $\theta = (\mu, \Lambda)$ have the distribution given by (2.17) and let $u = (u_1, \ldots, u_k) \mid \theta \sim N(\mu, \Lambda)$. Then the marginal distribution of $u$ is given by

\[
f_k^N(u) = \frac{(|\Psi_k|)^{1/2}}{(2\pi)^{kd/2}|\Psi_0|^{1/2}|\Sigma_0|^{1/2}} \exp \left(\frac{1}{2} W(u)\right)^{-1/2},
\]

(2.19)

where $\Gamma_d$ is the multivariate Gamma function and

\[
\Psi_k = (\Psi_0^{-1} + k \Sigma_0^{-1})^{-1}
\]

(2.20)

and

\[
W(u) = \sum_{i} (u_i - \mu_0)^t \Sigma_0^{-1} (u_i - \mu_0) - k^2 (\overline{u} - \mu_0)^t \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} (\overline{u} - \mu_0).
\]

(2.21)

**Proof.** Note that it is enough to prove (2.5) only for $\mu_0 = \overline{0}$; in general it follows by translation. The case $\mu_0 = \overline{0}$ is contained in Rajkowski [2018] as a part of the proof of Remark 2.1. \qed
2.2.3 Normal-inverse-Gamma

\[ \lambda \sim \mathcal{G}^{-1}(\beta_0 + 1, \beta_0) \]
\[ \Lambda = \lambda \Sigma_0 \]
\[ \mu | \lambda \sim \mathcal{N}(\mu_0, \lambda \Psi_0) \]

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

\[ \mathbb{E} \Lambda = \Sigma_0, \]
\[ \mathbb{V} \mu = \mathbb{E} \mathbb{V} \mu | \lambda + \mathbb{E} \mathbb{E} \mu | \lambda = \Sigma_0 + \lambda \Psi_0 = \Psi_0. \]

(2.22)

Proposition 2.6. Let \( \theta = (\mu, \Lambda) \) have the distribution given by (2.22) and let \( u = (u_1, \ldots, u_k) | \theta \sim^d \mathcal{N}(\mu, \Lambda) \). Then the marginal distribution of \( u \) is given by

\[ f_{\text{NIG}}(u) = \frac{\beta_0^{\alpha_k} |\Psi_k|^{1/2} \Gamma(\alpha_k)}{(2\pi)^{dk/2} |\Psi_0|^{1/2} \Gamma(\alpha_0)} \left( \beta(u) \right)^{-\alpha_k/2} \]

(2.24)

where \( \Psi_k \) is defined by (2.20),

\[ \alpha_k = \beta_0 + 1 + kd/2 \]

and

\[ \beta(u) = \beta_0 + \sum_i (u_i - \bar{u})^t \Sigma_0^{-1} (u_i - \bar{u}) + (\bar{u} - \mu_0)^t k \Xi_k (\bar{u} - \mu_0) \]

(2.25)

where

\[ \Xi_k = (\Sigma_0 + k \Psi_0)^{-1}. \]

(2.26)

Proof. The proof is left for Section 5.

2.2.4 Comparison of the models

If we assume that the component parameters in the Normal Bayesian Mixture Model are distributed by (2.17) then we assume that the covariance matrix in each component is equal to \( \Sigma_0 \) which is known to us. The results of Rajkowski [2018] imply that the misspecification of this hyperparameter may lead to serious inference issues regarding the number of clusters, at least as far as the MAP partition is concerned. In the light of these findings, (2.12) seems to be a safer choice of the prior for the component parameters. In this case the covariance matrix is chosen independently for each component according to the inverse-Wishart distribution. Note that although the Normal-inverse-Wishart prior gives more flexibility in terms of the component covariances, it imposes some modelling restriction, namely the expected within and between group covariance matrices are proportional, as is shown by (2.13). This does not affect the fixed-covariance model, cf. (2.18).

This is the reason for which we propose the Normal-inverse-Gamma prior. It is not listed in Gelman et al. [2013] and we were not able to find any reference to it in the literature. 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. At the same time, by allowing the component covariance matrix to scale between clusters can be a remedy to the drawbacks of fixed covariance prior that were pointed out in Rajkowski [2018].

As a final point we note that Normal-inverse-Gamma prior is a generalisation of the Normal prior in the sense that (2.22) becomes (2.17) as \( \beta_0 \to \infty \). Analogously, Normal-inverse-Wishart prior is a quasi-extension of the Normal prior, since as \( \eta_0 \to \infty \), (2.12) converges to (2.17), but with \( \Psi_0 = \Sigma_0/\kappa_0 \).
3 Formal statement of the result

We now define what we mean by linear, quadratic and elliptic separation of clusters.

Definition 3.1. Let \( \mathcal{X} \) be a family of subsets of \( \mathbb{R}^d \) and \( \mathcal{L} \) a family of real functions on \( \mathbb{R}^d \). We say that \( \mathcal{X} \) is separated by \( \mathcal{L} \) if for every \( X, Y \in \mathcal{X}, X \neq Y \), there exist \( L_{X,Y} \in \mathcal{L} \) such that \( L_{X,Y}(x) \geq 0 \) and \( L_{X,Y}(y) < 0 \) for all \( x \in X, y \in Y \). Moreover

- if \( \mathcal{L} \) are affine functions, i.e. \( \mathcal{L} = \{a^t x + b \text{ for some } a, b \in \mathbb{R}^d\} \), then \( \mathcal{X} \) is linearly separable.
- if \( \mathcal{L} \) are quadratic functions, i.e. \( \mathcal{L} = \{x^t A x + a^t x + b \text{ for some } A \in \mathbb{R}^{d \times d}, a, b \in \mathbb{R}^d\} \), then \( \mathcal{X} \) is quadratically separable.
- if \( \mathcal{L} \) is the union of convex and concave quadratic functions, i.e. \( \mathcal{L} = \{\pm x^t A x + ax + b \text{ for some } A \in \mathbb{S}^+, a, b \in \mathbb{R}^d\} \), then \( \mathcal{X} \) is elliptically separable.

(a) This family is linearly separable.

(b) This family is elliptically, but not linearly separable.

(c) This family is not quadratically separable.

Figure 1: Illustration of the different types of separability.

Notation. For the notational convenience we will use the separability notions also with respect to the sets of sequences in \( \mathbb{R}^d \). For example, if \( x_1, \ldots, x_n \in \mathbb{R}^d \) and \( I, J \) are disjoint subsets of \([n]\) then the expression \( x_I \) is linearly separated from \( x_J \) means that \( \{\{x_i : i \in I\}, \{x_j : j \in J\}\} \) is linearly separated.

In Rajkowski [2018] it is proved that in the Normal Bayesian Mixture Model with Normal distribution on the component mean and fixed covariance matrix, when the prior on the space of partitions is the Chinese Restaurant Process, the convex hulls of the clusters in the MAP partition are disjoint. Equivalently, the MAP is linearly separable. The same is true for an arbitrary exchangeable random partition prior.

Theorem 3.2. Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) be pairwise distinct and let \( \hat{I} \) be the MAP partition of \( x_1, \ldots, x_n \) in the Normal Bayesian Mixture Model where the prior on component parameters is given by (2.17). Then the family \( \{x_I : I \in \hat{I}\} \) is linearly separable.

Proof. The proof for the general exchangeable random partition prior is analogous to the proof in the case of the Chinese Restaurant Process, which is contained in Rajkowski [2018]. We give it in Section 5 for completeness.

The main result of this paper is an analogue of the Theorem 3.2 for Normal Bayesian Mixture Model with Normal-inverse-Wishart prior on the component parameters.

Theorem 3.3. Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) be pairwise distinct and let \( \hat{I} \) be the MAP partition of \( x_1, \ldots, x_n \) in the Normal Bayesian Mixture Model where the prior on component parameters is given by (2.12). Then the family \( \{x_I : I \in \hat{I}\} \) is quadratically separable.
Proof. The proof is left for Section 5.

By applying the same techniques as in the proof of Theorem 3.3 we can show that the MAP partition in the Normal Bayesian Mixture Model with Normal-inverse-Gamma prior on the component parameters is elliptically separable.

**Theorem 3.4.** Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) be pairwise distinct and let \( \hat{I} \) be the MAP partition of \( x_1, \ldots, x_n \) in the Normal Bayesian Mixture Model where the prior on component parameters is given by (2.22). Then the family \( \{ x_I : I \in \hat{I} \} \) is elliptically separable.

Proof. The proof is left for Section 5.

## 4 Discussion of potential applications

We proved linear or quadratic separability of the MAP partition in most popular Normal Bayesian Mixture Models. Apart from an aesthetic analogy to the properties of Fisher Discriminant Analysis, the benefits of such result may be twofold.

In Rajkowski [2018] the linear separability of the MAP partition is crucial for establishing the existence of ‘limits’ of the MAP partitions when the prior on partitions is the Chinese Restaurant Process and the data is independently and identically distributed with some ‘input distribution’. The limit is related to the partitions of observation space which maximises a given functional \( \Delta \) (which depends only on the hyperparameter \( \Sigma_0 \) and the input distribution). The linear separability is important for two reasons: firstly, it is possible to consider the limits of sequences of convex sets and secondly: it is possible to apply the Uniform Law of Large Numbers for the family of convex sets. Theorem 3.3 should enable an analogous approach for the Normal-inverse-Wishart and Normal-inverse-Gamma priors on the component parameters.

The other kind of application is more practical; Theorem 3.3 shows that the search for an MAP partition may be restricted to situations where clusters are quadratically separated. The space of such partitions is still far too large for an exhaustive search, but may help in finding a partition whose score approximates the MAP score.

## 5 Proofs

### 5.1 Proof of Proposition 2.6

Consider the model of Proposition 2.6

\[
\lambda \sim G^{-1}(\alpha_0, \beta_0) \propto \lambda^{\alpha_0-1} \exp \left\{ -\frac{\beta_0}{\lambda} \right\}
\]

\[
\mu | \lambda \sim N(\mu_0, \lambda \Sigma_0) \propto \lambda^{-d/2} \exp \left\{ -\frac{(\mu - \mu_0)^t (\lambda \Sigma_0)^{-1} (\mu - \mu_0)}{2} \right\}
\]

\[
u = (u_1, \ldots, u_k) | \mu, \lambda \iid N(\mu, \lambda \Sigma_0) \propto \lambda^{-kd/2} \exp \left\{ -\sum_{i=1}^k (u_i - \mu)^t (\lambda \Sigma_0)^{-1} (u_i - \mu) \right\}^{1/2},
\]

where \( \alpha_0 \) is defined by (2.25). The conditional density of \( \lambda, \mu \) given \( \nu \) is

\[
\lambda, \mu | \nu \propto \lambda^{\alpha_0-1-(k+1)d/2} \exp \left\{ -\lambda^{-1} \left( \beta_0 + (\mu - \mu_0)^t \Psi_k^{-1} (\mu - \mu_0) + \sum_{i=1}^k (u_i - \mu)^t \Sigma_0^{-1} (u_i - \mu) \right) \right\}^{1/2}.
\]

Recall the definition of \( \Psi_k (2.20) \). This is the update of the cluster covariance. Let

\[
\mu_k = \Psi_k \left( \Psi_k^{-1} \mu_0 + \sum_{i=1}^k u_i \right).
\]

(5.3)
This is the update of the cluster centre given the data. To get the part inside the exponential in a useful format:

\[
(\mu - \mu_0)^T \Psi_0^{-1} (\mu - \mu_0) + \sum_{i=1}^k (u_i - \mu)^T \Sigma_0^{-1} (u_i - \mu) = \\
= \sum_{i=1}^k u_i^T \Sigma_0^{-1} u_i + \mu_0^T \Psi_0^{-1} \mu_0 - 2\mu^T \left( \Psi_0^{-1} \mu_0 + \Sigma_0^{-1} \sum_{i=1}^k u_i \right) + \mu^T (\Psi_0^{-1} + k \Sigma_0^{-1}) \mu = \\
= \sum_{i=1}^k u_i^T \Sigma_0^{-1} u_i + \mu_0^T \Psi_0^{-1} \mu_0 - 2\mu^T \Psi_k^{-1} \mu_k + \mu^T \Psi_k^{-1} \mu = \\
= \sum_{i=1}^k u_i^T \Sigma_0^{-1} u_i + \mu_0^T \Psi_0^{-1} \mu_0 - \mu_k^T \Psi_k^{-1} \mu_k + (\mu - \mu_k)^T \Psi_k^{-1} (\mu - \mu_k).
\]

(5.4)

Therefore, setting \( \beta_k = \beta_0 + \sum_{i=1}^k u_i^T \Sigma_0^{-1} u_i + \mu_0^T \Psi_0^{-1} \mu_0 - \mu_k^T \Psi_k^{-1} \mu_k \):

\[
\lambda, \mu | \mathbf{u} \sim \mathcal{N}(\alpha_k, \beta_k) \\
\lambda, \mu | \mathbf{u} \sim \lambda^{-\alpha_k - 1 - d/2} \exp \left\{ - \lambda^{-1} (\beta_k + (\mu - \mu_k)^T \Psi_k^{-1} (\mu - \mu_k)) \right\}^{1/2}.
\]

(5.5)

Recall the definition of \( \Xi_k \) (2.27). Note that

\[
\Xi_k = \Psi_0^{-1} \Psi_k \Sigma_0^{-1} = \Sigma_0^{-1} \Psi_k \Psi_0^{-1}.
\]

(5.6)

Let us write

\[
\beta_k = \beta_0 + \sum_{i=1}^k u_i^T \Sigma_0^{-1} u_i + \mu_0^T \Psi_0^{-1} \mu_0 - \mu_k^T \Psi_k^{-1} \mu_k = \\
= \beta_0 + \sum_{i=1}^k (u_i - \bar{u})^T \Sigma_0^{-1} (u_i - \bar{u}) + k \bar{u}^T \Sigma_0^{-1} \bar{u} + \mu_0^T \Psi_0^{-1} \mu_0 - \mu_k^T \Psi_k^{-1} \mu_k = \\
= \beta_0 + \sum_{i=1}^k (u_i - \bar{u})^T \Sigma_0^{-1} (u_i - \bar{u}) + k \bar{u}^T \Sigma_0^{-1} \bar{u} + \mu_0^T \Psi_0^{-1} \mu_0 - \mu_k^T \Psi_k^{-1} \mu_k \\
- (\mu_0 \Psi_0^{-1} \Psi_k \Psi_0^{-1} \mu_0 + 2k \mu_0^T \Psi_0^{-1} \Psi_k \Sigma_0^{-1} \bar{u} + k^2 \bar{u}^T \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} \bar{u}) = \\
= \beta_0 + \sum_{i=1}^k (u_i - \bar{u})^T \Sigma_0^{-1} (u_i - \bar{u}) + k \bar{u}^T \Sigma_0^{-1} \Psi_k \Psi_0^{-1} \bar{u} + \mu_0^T \Psi_0^{-1} \Psi_k \Psi_0^{-1} \mu_0 - \\
- (\mu_0 \Psi_0^{-1} \Psi_k \Psi_0^{-1} \mu_0 + 2k \mu_0 \Xi_k \bar{u} + k^2 \bar{u}^T \Sigma_0^{-1} \Psi_k \Sigma_0^{-1} \bar{u}) \overset{(5.6)}{=} \\
= \beta_0 + \sum_{i=1}^k (u_i - \bar{u})^T \Sigma_0^{-1} (u_i - \bar{u}) + (k \mu_0 \Xi_k \mu_0 - 2k \mu_0 \Xi_k \bar{u} + k \bar{u}^T \Xi_k \bar{u}) = \\
= \beta_0 + \sum_{i=1}^k (u_i - \bar{u})^T \Sigma_0^{-1} (u_i - \bar{u}) + k(\mu_0 - \mu_k)^T \Xi_k (\mu_0 - \mu_k) = \beta(\mathbf{u}),
\]

where \( \beta(\cdot) \) is given by (2.26). Note that by (5.7) we have \( \beta_k \geq \beta_0 > 0 \) and hence by (5.5):

\[
\lambda | \mathbf{u} \sim \mathcal{G}^{-1}(\alpha_k, \beta_k) \\
\mu | \lambda, \mathbf{u} \sim \mathcal{N}(\mu_k, \Psi_k).
\]

(5.8)
We can now compute the marginal density of \( u \). It is the density of \((\Lambda, \mu, u)\) divided by the conditional density of \((\Lambda, \mu \mid u)\). We have
\[
\lambda, \mu, u \sim \beta_{\alpha_0}^{\beta_0} \Gamma^{-1}(\alpha_0) \lambda^{-\alpha_0 -1} - (k+1)d/2 |\Psi_0|^{-1/2} |\Sigma_0|^{-d/2} (2\pi)^{- (k+1)d/2}.
\]
\[
\lambda, \mu \mid u \sim \beta_k^{\beta_k} \Gamma^{-1}(\alpha_k) \lambda^{-\alpha_k -1 - d/2} |\Psi_k|^{-1/2} (2\pi)^{-d/2},
\]
\[
\cdot \exp \left\{ -\lambda^{-1} \left( \beta_k + (\mu - \mu_0)^\beta_k^{-1} (\mu - \mu_0) + \sum_{i=1}^k (u_i - \mu_i)^\beta_k^{-1} (u_i - \mu_i) \right) \right\}^{1/2},
\]
\[
\lambda, \mu \mid u \sim \beta_k^{\beta_k} \Gamma^{-1}(\alpha_k) \lambda^{-\alpha_k -1 - d/2} |\Psi_k|^{-1/2} (2\pi)^{-d/2}.
\]
\[
\cdot \exp \left\{ -\lambda^{-1} \left( \beta_k + (\mu - \mu_0)^\beta_k^{-1} (\mu - \mu_0) \right) \right\}^{1/2}.
\]
By definition of \( \Psi_k, \mu_k, \beta_k \) in the quotient of (5.9) and (5.10) the exponent function cancels out, leaving
\[
u \sim \frac{\beta_{\alpha_0}^{\beta_0} \Gamma(\alpha_0)}{\beta_k^{\beta_k} \Gamma(\alpha_k)} |\Psi_k|^{1/2} |\Sigma_0|^{-d/2} (2\pi)^{-kd/2} = \frac{\beta_{\alpha_0}^{\beta_0} |\Psi_k|^{1/2} \Gamma(\alpha_k)}{(2\pi)^{d/2} |\Psi_0|^{1/2} |\Sigma_0|^{d/2} \Gamma(\alpha_0)} \cdot \beta_k^{-\alpha_k/2}. 
\]
By (5.7) \( \beta_k = \beta(u) \) and the proof follows.

### 5.2 Two important Lemmas

**Lemma 5.1.** 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 \( \hat{I} \) be the MAP partition for \( x_1, \ldots, x_n \) in some Bayesian Mixture Model, given by (2.5). If for any \( U \subset [n], k, l \in \mathbb{N} \) such that \( k + l \neq |U| \) and \( I_{k,l} \) such that
\[
I_{k,l} \in \text{argmax}_{I \subset U : |I| = k} \left( \ln f_k(x_I) + \ln f_l(x_{U \setminus I}) \right)
\]
observations \( x_{I_{k,l}} \) and \( x_{U \setminus I_{k,l}} \) are separated by \( \mathcal{L} \) then \( \{x_I : I \in \hat{I}\} \) is separated by \( \mathcal{L} \).

**Proof.** Firstly note that by (2.11)
\[
\hat{I} \in \text{argmax}_{\text{partitions } I \subset [n]} \left( \ln p_n(I) + \sum_{I \in \hat{I}} \ln f_I(x_I) \right).
\]
Assume that the assumptions of Lemma 5.1 hold. Suppose that \( \hat{I} \) is not separated by \( \mathcal{L} \). Then there exist \( I, J \in \hat{I} \) such that \( x_I \) and \( x_J \) are not separated by \( \mathcal{L} \). Let \( U = I \cup J \) and \( k = |I| \). Let \( I = I_{k,l} \) and \( J \) be \( U \setminus I \). Moreover let \( \tilde{I} \) be a partition of \( [n] \) obtained by replacing \( I, J \) by \( I, J \), i.e. \( \tilde{I} = \hat{I} \setminus \{I, J\} \cup \{I, J\} \). Note that \( p_{\tilde{I}}(I) = p_n(I) \) (we have \( |I| = |\tilde{I}| \) and \( |J| = |\tilde{I}| \), so we use the exchangeability of \( \mathcal{L} \)). Moreover \( x_{\tilde{I}} \) and \( x_J \) are not separated by \( \mathcal{L} \) so by the assumptions of Lemma 5.1
\[
I \not\in \text{argmax}_{I \subset U : |I| = k} \left( \ln f_k(x_I) + \ln f_l(x_{U \setminus I}) \right)
\]
and hence
\[
\ln f_k(x_I) + \ln f_j(x_J) > \ln f_k(x_J) + \ln f_l(x_J).
\]
This means that
\[
\ln p_n(\tilde{I}) + \sum_{I \in \hat{I}} \ln f_I(x_I) > \ln p_n(\tilde{I}) + \sum_{I \in \hat{I}} \ln f_I(x_I),
\]
which contradicts the definition of \( \hat{I} \) and the proof follows. \( \square \)
Lemma 5.2. Let $V \subseteq \mathbb{R}^D$ be a convex set. Let $h: V \to \mathbb{R}$ be a strictly concave function, $z_1, \ldots, z_{k+1} \in \mathbb{R}^D$ are pairwise distinct. If $\sum_{i \in I} z_i \in V$ for every $I \subseteq [k+1]$ such that $|I| = k$ and

$$J \in \arg\min_{I \subseteq [k+1]: |I| = k} h\left(\sum_{i \in I} z_i\right)$$ (5.17)

then $z_J$ and $z_{[k+1] \setminus J}$ are linearly separable.

Proof. Consider the set of all possible sums of $k$ distinct vectors $z_i$, i.e. $S_k = \{\sum_{i \in I} z_i : I \subset [n], |I| = k\}$, and let $\hat{s}_k \in \arg\min_{s \in S_k} h(s)$. Since $h$ is strictly concave, then $\hat{s}_k$ is a vertex of conv $S_k$. This means that there exist a vector $v_0 \in \mathbb{R}^d$ such that $\hat{s}_k \in \arg\max_{s \in S_k} \langle s, v_0 \rangle$ (cf. Moszynska [2006], Corollary 3.3.6), where $\langle \cdot, \cdot \rangle$ is a standard Euclidean scalar product. I can also choose $v_0$ so that $\langle \hat{s}_i, v_0 \rangle$ are all different (because we are dealing with a discrete set). Let $z_{(1)}, \ldots, z_{(k+1)}$ be a decreasing ordering of vectors $z_i$ in the direction $v_0$’, i.e. $\{z_{(1)}, \ldots, z_{(k+1)}\} = \{z_1, \ldots, z_{k+1}\}$ and $\langle z_{(i)}, v_0 \rangle > \langle z_{(j)}, v_0 \rangle$ if $i < j$. Note that

$$\left\langle \sum_{i \in I} z_i, v_0 \right\rangle = \sum_{i \in I} \langle z_i, v_0 \rangle$$ (5.18)

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

5.3 Proof of Theorem 3.2

The proof of Theorem 3.2 when the prior distribution on the space of partitions is the Chinese Restaurant Process is contained in Rajkowski [2018]: the proof for general exchangeable random partitions does not differ from it. On the other hand, it is a direct consequence of Lemma 5.1 and Lemma 5.2, so we present it here for the completeness.

Let $U \subseteq [n]$, $k, l \in \mathbb{N}$ and $I_{k, l}$ be as in Lemma 5.1. Let $S = \sum_{i \in I} x_i$. Plugging the formula for $f_k^w$ (2.19) into (5.12) gives:

$$I_{k, l} \in \arg\max_{I \subseteq U : |I| = k} \left[k^2 \left(\frac{x_I}{k} - \mu \right)^T \Sigma_0^{-1} \Psi k \Sigma_0^{-1} \left(\frac{x_I}{k} - \mu \right) + l^2 \left(\frac{x_{I\setminus l}}{l} - \mu \right)^T \Sigma_0^{-1} \Psi l \Sigma_0^{-1} \left(\frac{x_{I\setminus l}}{l} - \mu \right)\right] = \arg\max_{I \subseteq U : |I| = k} w\left(\sum_{i \in I} x_i\right),$$ (5.19)

where

$$w(z) = k^2 \left(\frac{z}{k} - \mu \right)^T \Sigma_0^{-1} \Psi k \Sigma_0^{-1} \left(\frac{z}{k} - \mu \right) + l^2 \left(\frac{S - z}{l} - \mu \right)^T \Sigma_0^{-1} \Psi l \Sigma_0^{-1} \left(\frac{S - z}{l} - \mu \right).$$ (5.20)

Clearly, $w$ is a strictly convex quadratic function and hence by Lemma 5.2 $x_{I_{k, l}}$ and $x_{I_{l,k}}$ are linearly separable. By Lemma 5.1 this concludes the proof. \qed

5.4 Proof of Theorem 3.3

5.4.1 Löwner partial order, matrix convexity and concavity

Let $S^+$ be the set of positive definite matrices. Note that $\det A > 0$ for all $A \in S^+$ and hence log det is a well defined function on $S^+$.

Lemma 5.3. (Horn et al. [1990], Theorem 7.6.6) The function $f(A) = \ln \det A$ is a strictly concave function on $S^+$. [\[\text{(5.12)}\]\]}
Definition 5.4. The L"owner partial order is a partial order on the space of matrices where \( A \succeq B \) iff \( A - B \) is nonnegative definite (cf. Horn et al. [1990], Section 7.7). If \( A - B \) is positive definite, we write \( A \succ B \).

Definition 5.5. Let \( \varphi : \mathbb{R}^{d \times d} \to \mathbb{R} \) be a function. We say that it is increasing with respect to the L"owner partial order if \( \varphi(A) \geq \varphi(B) \) for all \( A \succeq B \). It is strictly increasing if \( \varphi(A) > \varphi(B) \) for all \( A \succeq B, A \neq B \).

Lemma 5.6. The determinant is strictly increasing on \( S^+ \) with respect to the L"owner partial order.

Proof. For \( A \in S^+ \) let \( \lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_d(A) \) be its eigenvalues. Let \( A \succeq B \geq 0 \) and \( A \neq B \). Horn et al. [1990], Corollary 7.7.4 (c) states that \( \lambda_i(A) \geq \lambda_i(B) \geq 0 \) for \( i \leq d \). Since the determinant is the product of eigenvalues it is enough to show that \( \lambda_i(A) > \lambda_i(B) \) for some \( i \leq d \). Suppose the contrary, i.e. \( \lambda_i(A) = \lambda_i(B) =: \lambda_i \) for all \( i \leq d \). Assume there are \( r \) distinct eigenvalues \( \lambda_1 > \ldots > \lambda_r \). Let \( V_0 = W_0 = \{0\} \) and let \( V_i, W_i \) be the eigenspace of \( \lambda_i \) in \( A \) and \( B \) respectively. The dimensions of \( V_i \) and \( W_i \) are equal since they are both equal to the number of repetitions of \( \lambda_i \) in the sequence \( \lambda_1, \ldots, \lambda_n \). By the Rayleigh Theorem (Horn et al. [1990], Theorem 4.2.2) for \( 1 \leq i \leq r \)

\[
V_i \setminus V_0 = \arg \max_{0 \neq u \in (\mathbb{R}^n)^* \setminus V_j} \frac{u^t Au}{u^t u} \quad \text{and} \quad W_i \setminus W_0 = \arg \max_{0 \neq u \in (\mathbb{R}^n)^* \setminus W_j} \frac{u^t Bu}{u^t u},
\]

where \( X^\perp \) is the space perpendicular to \( X \). We prove by induction that \( V_i = W_i \) for \( i \leq r \). The case \( i = 0 \) is clear. Assume that it is true for \( j \leq (i-1) \). Take any \( 0 \neq w \in W_i \). Using (5.21), the inequality \( A \succeq B \) and the assumption \( V_j = W_j \) for \( j \leq (i-1) \)

\[
\tilde{\lambda}_i = \frac{w^t B w}{w^t w} \leq \frac{w^t A w}{w^t w} \leq \max_{0 \neq u \in (\mathbb{R}^n)^* \setminus W_j} \frac{u^t A u}{u^t u} = \max_{0 \neq u \in (\mathbb{R}^n)^* \setminus W_j} \frac{u^t A u}{u^t u} = \tilde{\lambda}_i
\]

and hence \( \frac{w^t A w}{w^t w} = \tilde{\lambda}_i \) and \( w \in V_i \). This proofs that \( W_i \subset V_i \) and since \( \dim W_i = \dim V_i \) we get \( W_i = V_i \) for \( i = 1, 2, \ldots, r \). This means that \( A = B \), which is a contradiction. \( \square \)

Corollary 5.7. The function \( f(A) = \ln \det A \) is strictly increasing on \( S^+ \) with respect to the L"owner partial order.

Proof. The proof follows from Lemma 5.6. \( \square \)

Definition 5.8. Let \( \phi : \mathbb{R}^D \to \mathbb{R}^{d \times d} \) be a function. We say that \( \phi \) is matrix-convex if for all \( z_1, z_2 \in \mathbb{R}^D \) and \( p, q > 0 \) with \( p + q = 1 \)

\[
\phi(pz_1 + qz_2) \preceq p\phi(z_1) + q\phi(z_2).
\]

If the equality holds if and only if \( z_1 = z_2 \) we say that \( \phi \) is strictly matrix-convex. If \( -\phi \) is (strictly) matrix-convex then \( \phi \) is (strictly) matrix-concave.

Lemma 5.9. Let \( \phi : \mathbb{R}^d \to \mathbb{R}^{d \times d} \) be given by \( \phi(u) = uu^t \). Then \( \phi \) is strictly matrix-convex.

Proof. Let \( p, q > 0, p + q = 1 \). The proof follows from the following equality

\[
(pz_1 + qz_2)(pz_1 + qz_2)^t + p(z_1 - z_2)(z_1 - z_2)^t = pz_1 z_1^t + qz_2 z_2^t
\]

using \( pq > 0, (z_1 - z_2)(z_1 - z_2)^t \geq 0 \) and \( (z_1 - z_2)(z_1 - z_2)^t = 0 \iff z_1 = z_2 \). \( \square \)
5.4.2 Theorem 3.3

For \( I \subseteq [n] \) let

\[
\Sigma_I = \Sigma(x_I),
\]

where \( \Sigma(\cdot) \) is defined by (2.16). Let \( U \subseteq [n] \), \( k, l \in \mathbb{N} \) and \( I_{k,l,U} \) be as in Lemma 5.1. Plugging the formula for \( f_{k,l}^{\text{opt}} \) (2.14) into (5.12) gives:

\[
I_{k,l,U} = \arg\min_{I \subseteq U, |I| = k} (\nu_k \ln \det \Sigma_I + \nu_l \ln \det \Sigma_{U\setminus I}).
\]

(5.26)

Take any \( I \subseteq [n], |I| = k \). We have to show that \( \Sigma_I \) can be expressed in terms of base functions. To this end

\[
\Sigma_I = \Sigma_0 + \sum_{I} x_i x_i' - k x_ix_i' + \frac{k_0}{k_k} (\mu_0 - \mu_0'^t) x_i' x_i + \frac{k_0}{k_k} \mu_0 \mu_0' + \frac{k_0}{k_k} \mu_0' \mu_0 =
\]

(5.27)

\[
= \Sigma_0 + \sum_{I} x_i x_i' - k x_ix_i' + \frac{k_0}{k_k} \mu_0 \mu_0' + \frac{k_0}{k_k} \mu_0' \mu_0 + \frac{k_0}{k_k} \mu_0 \mu_0' =
\]

\[
= \Sigma_0 + \sum_{I} x_i x_i' - \frac{1}{k_k} \sum_i x_i (\sum_i x_i)' - \frac{k_0}{k_k} \mu_0 (\sum_i x_i)' - \frac{k_0}{k_k} (\sum_i x_i)' \mu_0 + \frac{k_0}{k_k} \mu_0 \mu_0' + \frac{k_0}{k_k} \mu_0 \mu_0'.
\]

For \( x \in \mathbb{R}^d, x = (x_1, \ldots, x_d) \)

\[
\Sigma = \begin{pmatrix} x_1^2 & x_1 x_2 & \cdots & x_1 x_d \\ x_2 x_1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_d x_1 & \cdots & x_d x_{d-1} & x_d^2 \end{pmatrix}
\]

and let \( D = d(d+1) \). Note, that \( \Sigma_I = \sigma_1 \left( \sum_I \Sigma \right) \), where for \( z \in \mathbb{R}^D, z = (z_1, \ldots, z_{|D|}) \):

\[
\sigma_1(z) = \Sigma_0 + \left[ z_{((i-1)+j+1)} \right]_{i,j \leq d} - \frac{1}{k_k} z_{((D-d+1):D)} z_{((D-d+1):D)}^t - \frac{k_0}{k_k} z_{((D-d+1):D)} z_{((D-d+1):D)}^t + \frac{k_0}{k_k} \mu_0 \mu_0'.
\]

(5.29)

Note that the only nonlinear part of \( \sigma_1 \) is \( -\frac{1}{k_k} z_{((D-d+1):D)} z_{((D-d+1):D)}^t \). By Lemma 5.9 it follows that \( \sigma_1 \) is strictly matrix-concave.

Let \( V_1 = \sigma_1^{-1}(S^+) \subseteq \mathbb{R}^D \). By the matrix-concavity of \( \sigma_1 \) we get that \( V_1 \) is a convex set. By (5.25) and (2.16) we also know that \( \Sigma_I \in S^+ \), so \( \sum_I \Sigma \in V_1 \), for all \( I \subseteq U \). Since \( \sigma_1 \) is strictly matrix-concave and log det is strictly increasing with respect to the Löwner partial order (Corollary 5.7) and a concave function on \( S^+ \) (Lemma 5.3), we have for any \( p, q > 0 \), \( p + q = 1 \) and \( z_1, z_2 \in V_1 \):

\[
\ln \det \sigma_1(p z_1 + q z_2) > \ln \det (p \sigma_1(z_1) + q \sigma_1(z_2)) \geq p \ln \det \sigma_1(z_1) + q \ln \det \sigma_1(z_2),
\]

(5.30)

Hence \( \ln \det \sigma_1(z) \) is a strictly concave function on \( V_1 \). Now let \( Q = \sum_I x_i x_i' \) and \( S = \sum_i x_i \).

Then \( \Sigma_{\alpha, I} = \sigma_2 \left( \sum_I \Sigma \right) \) where for \( z \in \mathbb{R}^d \):

\[
\sigma_2(z) = \Sigma_0 + \left( Q - z_{((i-1)+j+1)} \right)_{i,j \leq d} - \frac{1}{k_k} (S - z_{((D-d+1):D)}) (S - z_{((D-d+1):D)})^t - \frac{k_0}{k_k} \mu_0 (S - z_{((D-d+1):D)})^t + \frac{k_0}{k_k} \mu_0 \mu_0'.
\]

(5.31)
Let $V_2 = \sigma_2^{-1}(S^+)$. In the same way as before we can prove that $V_2$ is a convex set, $\sum_I \mathbf{z}_i \in V_2$ for all $I \subseteq \mathcal{U}$ and $\ln \det \sigma_i(z)$ is strictly concave on $V_2$. Hence we have that

$$\hat{I}_{k,\ell} = \arg\min_{I \subseteq \mathcal{U} : |I| = k} \left( \nu_k \ln \det \sigma_1 \left( \sum_{t \in I} \mathbf{z}_t \right) + \nu_l \ln \det \sigma_2 \left( \sum_{t \in I} \mathbf{z}_t \right) \right), \quad (5.32)$$

where $\nu_k \ln \det \sigma_1(\cdot) + \nu_l \ln \det \sigma_2(\cdot)$ is a strictly concave function on $V = V_1 \cap V_2$. Therefore by Lemma 5.2 we obtain that $\Sigma_{\hat{I}_{k,\ell}}$ and $\Sigma_{\hat{X}_{k,\ell}}$ are linearly separable (i.e. in terms of the base functions). Obviously this yields quadratic separability of $\mathbf{x}_{\hat{I}_{k,\ell}}$ and $\mathbf{x}_{\hat{U} \setminus \hat{I}_{k,\ell}}$ and the proof follows.

5.5 Proof of Theorem 3.4

This proof has strong similarities to the proof of Theorem 3.3, but the additional structure leads to sharper results, which require different details. We therefore present the proof in full.

For $I \subseteq [n]$ let

$$\beta_I := \beta(\mathbf{x}_I),$$

(5.33)

where $\beta(\cdot)$ is defined by (2.26). Let us make the following substitution: $y_i = \sigma_0^{-1/2}x_i$, and let $\tilde{\mu}_0 = \sigma_0^{-1/2}\tilde{\mathbf{z}}_k$ and $\tilde{\Sigma}_k = \sigma_0^{-1/2}\tilde{\mathbf{z}}_k\sigma_0^{-1/2}$. Then

$$\beta_I = \beta_0 + \sum_I (y_i - \bar{y}_I)^t(\bar{y}_I - \tilde{\Sigma}_k\bar{y}_I - \tilde{\mu}_0).$$

(5.34)

This variable change will be the key difference which will give the sharper separability result. Let $\mathcal{U} \subseteq [n]$, $k, l \in \mathbb{N}$ and $\hat{I}_{k,\ell}$ be as in Lemma 5.1. Plugging the formula for $f_k^{NIG}$ (2.24) into (5.12) gives:

$$\hat{I}_{k,\ell} \in \arg\min_{I \subseteq \mathcal{U} : |I| = k} \left( \alpha_k \ln \beta_I + \alpha_l \ln \beta_{\hat{U} \setminus I} \right).$$

(5.35)

We have to show that $\beta_I$ can be expressed in terms of the base functions. To this end

$$\beta_I = \beta_0 + \sum_I y_i^t(y_i - \bar{y}_I)^t(\bar{y}_I - \tilde{\Sigma}_k\bar{y}_I - \tilde{\mu}_0) =$$

$$= \beta_0 + \sum_I y_i^t(y_i - \bar{y}_I)^t(\bar{y}_I - \tilde{\Sigma}_k\bar{y}_I) - 2\tilde{\Sigma}_k\bar{y}_I + \tilde{\mu}_0^t\tilde{\Sigma}_k\tilde{\mu}_0 =$$

$$= \beta_0 + \sum_I y_i^t(y_i - \bar{y}_I)^t(I_d - \tilde{\Sigma}_k) + 2\tilde{\Sigma}_k\bar{y}_I + \tilde{\mu}_0^t\tilde{\Sigma}_k\tilde{\mu}_0.$$

(5.36)

For $y \in \mathbb{R}^d$, $y = (y(1), \ldots, y(d)) \in \mathbb{R}^{2d}$. Let

$$\mathbf{y} = (y(1), \ldots, y(d))^t \in \mathbb{R}^{2d}.$$  

(5.37)

Note that $\beta_I = b_1 \left( \sum \mathbf{y} \right)$, where for $z \in \mathbb{R}^{2d}$, $z = (z(1), \ldots, z(2d))$:

$$b_1(z) = \beta_0 + \sum_{j=1}^d z(j)^t z(j)^t(I_d - \tilde{\Sigma}_k)z(j) + \tilde{\mu}_0^t\tilde{\Sigma}_k z(j) + \tilde{\mu}_0^t\tilde{\Sigma}_k \tilde{\mu}_0.$$  

(5.38)

Now we show that $b_1(z)$ is a strictly concave function. It is a quadratic function and to prove its strict concavity it is enough to show that $(I_d - \tilde{\Sigma}_k)$ is positive definite, i.e. $I_d \succ \tilde{\Sigma}_k$ (cf. Definition 5.4). This is equivalent to $\tilde{\Sigma}_k^{-1} \succ I_d$ (cf. Horn et al. [1990], Corollary 7.7.4), which in turn is equivalent to $\Sigma_k^{-1} \succ \Sigma_0$. The latest is obvious as $\Sigma_k^{-1} = \Sigma_0 + \psi_0$. 

13
Let $V_1 = \mathbb{R}_+ \setminus \mathbb{R}^{2d}$. By the concavity of $b_i$ we get that $V_1$ is a convex set. By (5.33) and (2.26) we have $\beta_i \geq \beta_0 > 0$, so $\sum_{i} y_i \in V_1$, for all $I \subseteq \mathcal{U}$. Since $b_i$ is strictly concave and the logarithm is strictly increasing and concave, we have for any $p, q > 0$, $p + q = 1$ and $z_1, z_2 \in V_1$, $z_1 \neq z_2$:

$$\ln b_1(pz_1 + qz_2) > \ln (pz_1 + qz_2) \geq p \ln b_1(z_1) + q \ln b_1(z_2),$$

hence $\ln b_1(z)$ is a strictly concave function on $V_1$. Now let $Q = \sum_{i \in \mathcal{U}} y_i^l y_i$ and $S = \sum_{i \in \mathcal{U}} y_i$.

Then $\beta_{l|l} = b_2(\sum_i y_i)$ where for $z \in \mathbb{R}^{2d}$:

$$b_2(z) = \beta_0 + (Q - \sum_{j=1}^d z_{(j)}) - k^{-1}((S - z_{((d+1):(2d))})(I_d - \frac{\mu_k}{\sqrt{m}})(S - z_{((d+1):(2d)}) + 2\mu_k \sqrt{m} P_{(d+1):(2d)} + \mu_k \sigma^2 + \mu_k^2 \bar{\nu}_k \bar{\mu}_0. $$

Let $V_2 = \mathbb{R}_+ \setminus \mathbb{R}^{2d}$. In the same way as before we can prove that $V_2$ is a convex set, $\sum_i y_i \in V$ for all $I \subseteq \mathcal{U}$ and $\ln b_2(z)$ is strictly concave on $V_2$. Hence we have that

$$I_{l|l} = \arg\min_{I \subseteq \mathcal{U} : |I| = k} \left( \alpha_l \ln b_1 \left( \sum_{i \in I} y_i \right) + \alpha_i \ln b_2 \left( \sum_{i \in I} y_i \right) \right),$$

where $\alpha_l \ln b_1(\cdot) + \alpha_i \ln b_2(\cdot)$ is a strictly concave function on $V = V_1 \cap V_2$. Therefore if we apply Lemma 5.2 it follows that $x_{l|l}$ and $x_{I \setminus I_{l|l}}$ are linearly separable (i.e. in terms of the base functions). Obviously this yields elliptic separability of $y_{l|l}$ and $x_{I \setminus I_{l|l}}$, which is equivalent to the elliptic separability of $x_{l|l}$ and $x_{I \setminus I_{l|l}}$ and the proof follows. \qed

References

David J Aldous. Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII—1983, pages 1–198. Springer, 1985.

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The annals of statistics, pages 1152–1174, 1974.

David Blackwell, James B MacQueen, et al. Ferguson distributions via polyà urn schemes. The annals of statistics, 1(2):353–355, 1973.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. The annals of statistics, pages 209–230, 1973.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA:, 2001.

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.

Roger A Horn, Roger A Horn, and Charles R Johnson. Matrix analysis. Cambridge university press, 1990.

Maria Moszynska. Selected topics in convex geometry. Springer, 2006.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. def. 1(2σ2):16, 2007.

Jim Pitman. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course, 2002.
Lukasz Rajkowski. Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Analysis*, 2018.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Tatiana Tatarinova and Alan Schumitzky. *Nonlinear Mixture Models: A Bayesian Approach*. World Scientific, 2015.