Contributions to Bayesian model selection in finite and infinite mixtures with an application to distributed computing

Adrien Hairault

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL
Préparée à l'Université Paris-Dauphine

Contributions to Bayesian model selection in finite and infinite mixtures with an application to distributed computing

Soutenue par
Adrien HAIRAULT
Le 10 novembre 2023

Ecole doctorale n° ED 543
Ecole doctorale SDOSE

Spécialité
Mathématiques

Composition du jury :

Pierre, JACOB Professeur, ESSEC Business school	*Président*
François, CARON Professeur, University of Oxford	*Rapporteur*
Anne, PHILIPPE Professeur, Université de Nantes	*Rapporteur*
Robin, RYDER Maître de conférence, Université Paris-Dauphine PSL	*Examinateur*
Christian, ROBERT Professeur, Université Paris-Dauphine PSL	*Directeur de thèse*
Judith, ROUSSEAU Professeur, University of Oxford	*Directrice de thèse*

Dauphine | PSL
UNIVERSITÉ PARIS

À mes grands-parents.

Remerciements

En premier lieu, je tiens à remercier mes directeurs de thèse Judith Rousseau et Christian Robert. Votre accompagnement durant ces trois années et demi ponctuées de périodes de confinement a été crucial pour la réussite de ce projet. Je tiens à vous remercier d'avoir partagé votre expertise et votre temps. Je suis honoré d'avoir eu l'opportunité de travailler sous votre direction.

Je remercie les rapporteurs François Caron et Anne Philippe ainsi que les autres membres du jury Pierre Jacob et Robin Ryder pour leur relecture et leurs remarques sur le présent manuscrit.

Je remercie à nouveau mon directeur de thèse Christian Robert pour son appui financier dans le cadre de sa chaire PRAIRIE (ANR-19-P3IA-0001) m'ayant permis de réaliser ce doctorat dans d'excellentes conditions ainsi que de pouvoir assister à de nombreuses conférences. Je lui suis profondément reconnaissant de la confiance qu'il m'a accordée.

Il m'est impossible de ne pas mentionner l'ensemble des personnels administratifs du CEREMADE sans qui ces années de thèse auraient été autrement plus compliquées. Merci à César Faivre, Isabelle Bellier et Anne-Laure Chagnon pour leur travail et leur bienveillance. Un grand merci également à Gilles Bares et Thomas Duleu pour leur appui et leur efficacité dans la résolution de mes quelques problèmes avec le cluster. Merci à Vincent Rivoirard qui a dirigé le laboratoire tout au long de ma thèse. Merci enfin à l'Ecole doctorale et particulièrement à Béatrice de Tilière pour son aide précieuse.

Je remercie Donato, Ryad et Théo pour leur amitié et leur soutien indéfectible. Merci aux autres doctorants et particulièrement à Amirali, Antoine, Charly, Claudia, Clément, Emma, João, Lorenzo et Quan pour tous ces crous-cous engloutis et ces discussions partagées.

Merci à Olga, Marina, Félix, Alex et Lucas pour tous ces bons moments.

Merci à ma soeur Rebecca pour nos voyages et nos parties de crapette.

Merci à Adrien. Ces trois années auraient été bien moins belles si elles n'avaient pas été à tes côtés.

Je remercie également ma famille; ma mère pour son soutien moral en toute circonstance et ses excellents conseils, mon père pour ces vacances passées ensemble qui m'ont véritablement permis de m'aérer l'esprit. Mes grands-parents, à qui ce travail est dédié, pour leur gentillesse et leur amour. Vous rendre fiers est ma plus belle récompense.

Contents

Résumé

Estimation de la vraisemblance marginale des modèles de mélange finis

Les modèles de mélange suscitent un intérêt considérable en raison de leur capacité à modéliser l'hétérogénéité dans une population donnée. À cet égard, ils ont été introduits formellement pour la première fois dans Pearson 1894, où un mélange normal à deux composantes est ajusté au rapport entre la longueur du front et la longueur du corps d'une population de crabes. En établissant une correspondance entre moments empiriques et moments théoriques d'une loi gaussienne puis en résolvant une équation polynomiale de degré neuf, Pearson identifie deux groupes homogènes au sein de la population de crabes, mettant ainsi en évidence une probable divergence évolutive entre les deux groupes, comme illustré en Figure 1.

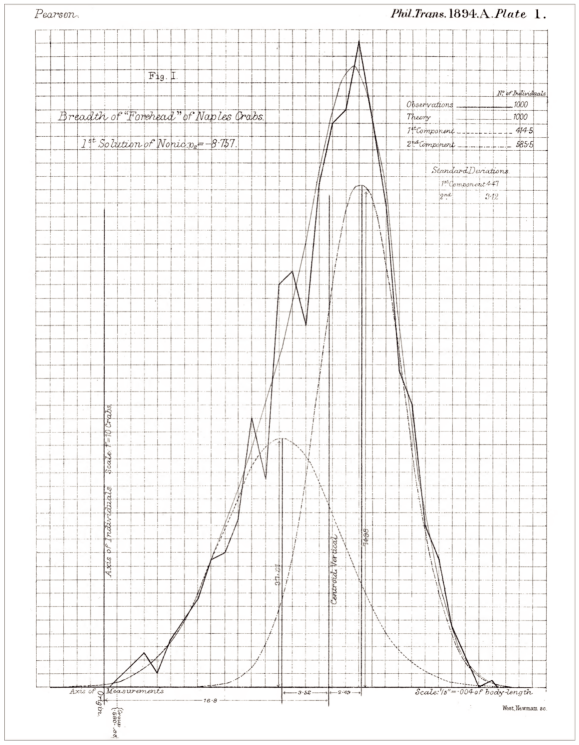


Figure 1: Le modèle de mélange à deux composantes ajusté dans Pearson 1894 à la population de crabes étudiée. La ligne continue est une *courbe de fréquence*. Les lignes en pointillés représentent les deux composantes pondérées du mélange de loi Normales.

Spécification du modèle et sélection de son ordre

Formellement, nous disons qu'un ensemble de données indépendantes et identiquement distribuées (*i.i.d.*) $\boldsymbol{y} = (y_1, \ldots, y_n)$ provient d'un modèle de mélange fini, en anglais Finite Mixture (FM), à K composantes si leur densité peut être écrite sous la forme

$$p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \varpi_k f(y_i|\theta_k)$$

pour $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi}) \in \Theta^K \times \Delta_{K-1}$ où $\Theta \subset \mathbb{R}^d$ et où nous désignons par $\Delta_{K-1} := \{\boldsymbol{\varpi} \in (0,1)^K : \sum_{k=1}^{K} \varpi_k = 1\}$ le simplexe de dimension $(K-1)$. Le vecteur $(\varpi_1, \ldots, \varpi_K)$ contient les poids du mélange tandis que les paramètres du mélange $(\theta_1, \ldots, \theta_K)$ paramétrisent une certaine densité f qui est généralement appelée *noyau du mélange*.

La modélisation par les mélanges peut en grande partie se réduire au choix d'un noyau de mélange approprié f et du bon ordre du modèle K, qui correspond au nombre de composantes du mélange. Ce dernier nécessite généralement une sélection et/ou une estimation soignée et revêt un intérêt particulier lorsque l'objectif principal de l'inférence est lié au regroupement (*clustering*). En effet, les modèles de mélange sont devenus l'un des principaux outils utilisés par qui souhaite identifier des sous-groupes avec des caractéristiques spécifiques au sein d'une population donnée (McLachlan and Basford 1988; Geary 1989; Dang et al. 2023). Des applications sim-

ilaires, mais plus complexes, comprennent la modélisation de sujets (*topic modeling*) tels que l'allocation Dirichlet latente (Blei et al. 2003; Chen and Doss 2019), ou la segmentation, la compression et la classification d'images (Aiyer et al. 2005; Zeng et al. 2014).

Les méthodes existantes pour estimer K sont présentées dans l'examen approfondi de Celeux et al. 2019. Elles comprennent des stratégies fréquentistes telles qu'une version adaptées des tests de rapport de vraisemblance (*Likelihood ratio test* ou LRT, McLachlan 1987; Heckman et al. 1990; McLachlan and Peel 2000; Frühwirth-Schnatter 2006), des estimateurs issus de la méthode des moments (Dacunha-Castelle and Gassiat 1997), ou bien des critères d'information populaires (Smyth 2000).

D'un point de vue bayésien, l'estimation de K peut être effectuée en même temps que l'estimation des paramètres et des poids de mélange $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi})$ en définissant une distribution a priori sur K. Échantillonner selon la distribution a posteriori de K est loin d'être simple et nécessite la conception d'échantillonneurs dits *transdimensionnels*. Les travaux fondateurs de Green 1995 et Richardson and Green 1997 ont conduit à la construction de méthodes de Monte-Carlo par chaînes de Markov à sauts réversibles (RJMCMC) qui permettent de réaliser de tels "sauts" entre des espaces de dimensions différentes. De tels modèles où K est traité comme n'importe quel autre paramètre ont été formalisés par Richardson and Green 1997 ou Nobile 2004, où ils étaient appelés mélanges avec un nombre aléatoire de composantes. Le terme *mélange de mélanges finis* (MFM, Miller and Harrison 2018) est maintenant plus fréquemment utilisé. De récentes avancées incluent notamment l'établissement de nouvelles stratégies d'échantillonnage adaptées à la complexité intrinsèque des mélanges (voir par exemple Frühwirth-Schnatter et al. 2021).

Une autre approche bayésienne consiste à calculer le facteur de Bayes (Jeffreys 1935; Raftery 1996) pour des valeurs concurrentes de K. Une telle stratégie est connue pour être consistante (c'est-à-dire que le facteur de Bayes pointe de manière cohérente vers le bon modèle pour un nombre croissant d'observations n, Chib and Kuffner 2016). En pratique, les applications concrètes de cette approche nécessitent l'estimation de la vraisemblance marginale d'un modèle de mélange fini, définie comme l'intégrale de la fonction de vraisemblance par rapport à la distribution a priori, ce qui n'est pas une tâche facile. Les algorithmes d'échantillonnage de Monte Carlo les plus populaires pour relever ce défi sont l'algorithme de Chib (Chib 1995), l'algorithme *bridge sampling* (Meng and Wong 1996; Frühwirth-Schnatter 2004; Frühwirth-Schnatter 2019) et, dans une moindre mesure, les méthodes de Monte Carlo séquentielles (Chopin 2002; Gunawan et al. 2020).

De manière plus générale, être en mesure de calculer la vraisemblance marginale d'un modèle revêt une importance cruciale dans un contexte bayésien car c'est là l'outil principal utilisé pour l'évaluation et la comparaison de modèles. En effet, Fong and Holmes 2020 montre que le calcul de la vraisemblance marginale d'un modèle est formellement équivalent à une validation croisée exhaustive de type *leave-p-out* moyennée sur toutes les valeurs de p, la validation croisée étant la procédure fréquentiste de référence pour l'évaluation de modèles. Par conséquent, être en mesure de calculer la vraisemblance marginale d'un modèle de mélange fini est d'intérêt non seulement pour sélectionner de manière cohérente une valeur de K,

mais également pour trouver un noyau de mélange approprié f, ou même pour comparer un modèle de mélange à une alternative paramétrique ou non paramétrique.

Contributions du Chapitre 2

Une approximation sensée de la vraisemblance marginale des modèles de mélange nécessite des méthodes *ad hoc* qui tiennent compte des défis spécifiques liés à ces modèles. En particulier, pour une distribution a priori echangeable sur les poids et les paramètres de mélange, la distribution a posteriori du mélange $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$ est invariante par permutation. Cela se traduit par

$$\pi((\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K)|\boldsymbol{y}) = \pi((\boldsymbol{\vartheta}_{\sigma(1)}, \ldots, \boldsymbol{\vartheta}_{\sigma(K)}|\boldsymbol{y})$$

pour tous $(\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K) \in \Theta^K \times \Delta_{K-1}$ et toutes les permutations $\sigma \in \mathfrak{S}_K$, où $\mathfrak{S}_K$ désigne l'ensemble des permutations de $\{1, \ldots, K\}$.
Cela implique que la distribution a posteriori présente $K!$ modes équiprobables. Comme l'a remarqué Neal 1999 en réponse à Chib 1995, un bon comportement de mélange des algorithmes MCMC ciblant la distribution a posteriori est essentiel pour une approximation satisfaisante de la vraisemblance marginale du modèle. Cependant, dans le cas des mélanges finis, cela nécessite une visite équilibrée de toutes les $K!$ configurations modales de la distribution a posteriori (un phénomène également appelé *label switching*), ce qui la plupart du temps est une attente irréaliste étant donné un budget computationnel fini. Pour compenser ce comportement indésirable des échantillonneurs MCMC, les méthodes populaires telles que l'algorithme de Chib et l'algorithme *bridge sampling* ainsi que leurs adaptations ultérieures aux mélanges finis ont recours à l'application artificielle d'un phénomène de permutation parfait en intégrant leurs estimateurs sur l'espace des permutations $\mathfrak{S}_K$. Cela entraîne un coût supplémentaire de $O(K!)$ qui n'est en pratique supportable que pour des valeurs de K inférieures 5. Dans le Chapitre 2, nous introduisons d'abord un algorithme de Chib modifié qui utilise la structure de partition induite par les modèles de mélange. Cette dernière a en effet la propriété attirante de résister au phénomène de *label switching*, évitant ainsi le coût exponentiel de $O(K!)$ payé par les méthodes traditionnelles. Nous adaptons également l'algorithme d'imputation séquentielle découvert par Kong et al. 1994 aux mélanges finis. En plus de sa robustesse vis-à-vis du label switching, cette approche se révèle également robuste à une augmentation du nombre d'observations n. Finalement, nous proposons une revue empirique des estimateurs classiques de la vraisemblance marginale, anciens et nouveaux, et mettons en évidence leurs forces et faiblesses dans différents scénarios où K et n ne sont pas nécessairement petits. En particulier, nous constatons que les approches de Monte Carlo séquentielles sont les plus efficaces et nous espérons que cette évaluation sera une incitation à utiliser ces méthodes plus souvent.

Propriétés asymptotiques et estimation de la vraisemblance marginale pour les modèles de mélange de processus de Dirichlet

Spécification du modèle

Le modèle de mélange de processus de Dirichlet (DPM), introduit pour la première fois dans Ferguson 1983, est l'un des principaux outils du domaine des statistiques bayésiennes non-paramétriques. En supposant a priori un nombre infini de composantes de mélange, il est parfois appelé modèle de mélange "infini" par opposition aux mélanges finis. En effet, pour des données $i.i.d$ $\boldsymbol{y} = (y_1, \dots, y_n)$, le DPM peut être décrit par la spécification du modèle génératif suivant :

$$\begin{aligned} y_i|\theta_i &\overset{i.i.d}{\sim} F(y_i|\theta_i) \quad \text{pour } i = 1, \dots, n \\ \theta_i|P &\sim P \\ P|M &\sim DP(M, G_0) \\ M &\sim \Pi_M \end{aligned}$$

où $F(\cdot|\theta)$ est une distribution à support dans $\mathbb{R}^d$ avec une densité par rapport à la mesure de Lebesgue $f(\cdot|\theta)$ et $DP(M, G_0)$ désigne le processus de Dirichlet ayant pour mesure de base G_0 et paramètre de concentration M. Les réalisations P du processus de Dirichlet sont presque sûrement discrètes, de sorte que, si δ_x désigne la fonction delta de Dirac,

$$P = \sum_{k=1}^{\infty} \varpi_k \delta_{\theta_k}$$

avec $\varpi_k = V_k \prod_{l<k}(1 - V_l)$ où $V_l \overset{i.i.d}{\sim} Beta(1, M)$ (assurant que $\sum_{k=1}^{\infty} \varpi_k = 1$ presque sûrement) et $\theta_1, \theta_2, \dots \overset{i.i.d}{\sim} G_0$.
La vraisemblance de P peut alors s'écrire comme

$$f_P(\boldsymbol{y}) := p(\boldsymbol{y}|P) = \prod_{i=1}^{n} \int_{\Theta} f(y_i|\theta) dP(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{\infty} \varpi_k f(y_i|\theta_k).$$

Le champ d'applications du DPM est vaste, allant du regroupement (clustering) multivarié (Crépet and Tressou 2011) à l'estimation bayésienne de densités (Neal 1992; Rabaoui et al. 2012). Une caractéristique attrayante du DPM est l'hypothèse a priori très faible posée sur la distribution des poids de mélange $\boldsymbol{\varpi} = (\varpi_1, \varpi_2, \dots)$ et des paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots)$ puisque leur distribution a priori P est elle-même aléatoire, ce qui en fait l'une des pierres angulaires de la statistique bayésienne non-paramétrique.

Consistance du facteur de Bayes pour le test d'une hypothèse nulle paramétrique par rapport à une alternative non-paramétrique de mélange de processus de Dirichlet

Recourir à ce type de modèles flexibles car non paramétriques peut être motivé par le souhait de se libérer de la contrainte de la sélection et de la comparaison de modèles. Par conséquent, il peut sembler surprenant au départ de s'intéresser à l'estimation de la vraisemblance marginale du DPM, définie par

$$m_{DP}(\boldsymbol{y}) = \int f_P(\boldsymbol{y})\Pi(dP, dM)$$

où Π désigne la distribution a priori conjointe sur (P, M) avec $P|M \sim DP(M, G_0)$ et $M \sim \Pi_M$.
Cependant, il arrive que la quantité $m_{DP}(\boldsymbol{y})$ puisse avoir plusieurs applications d'intérêt. Par exemple, le problème habituel de modélisation statistique qui consiste à déterminer si un échantillon de taille n $\boldsymbol{y} = (y_1, \ldots, y_n)$ provient d'une famille paramétrique particulière de distributions peut être formalisé comme un test d'adéquation par rapport à une alternative non paramétrique, comme le DPM. Formellement, nous considérons ici le problème de tester l'hypothèse nulle paramétrique $H_0 : f_0 \in \cup_{K\in\mathbb{N}^*}\mathfrak{M}_K$, où $\mathfrak{M}_K$ désigne le modèle paramétrique de mélanges de K distributions $f_\theta(y)$, par rapport à l'alternative $H_1 : f_0 \in \mathcal{F} \setminus \{\cup_{K\in\mathbb{N}^*}\mathfrak{M}_K\}$ modélisée par le modèle de mélange de processus de Dirichlet avec noyau de mélange $f_\theta(y)$, pour $\mathcal{F}$ un ensemble englobant de fonctions de densité. Cette procédure est pertinente du point de vue bayésien à condition que le facteur de Bayes soit consistant, c'est-à-dire si

$$\lim_{n\to\infty} BF_{0,1} := \lim_{n\to\infty} \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} = \begin{cases} \infty & \text{sous } H_0 : f_0 \in \cup_{K\in\mathbb{N}^*}\mathfrak{M}_K \\ 0 & \text{sous } H_1 : f_0 \in \mathcal{F} \setminus \{\cup_{K\in\mathbb{N}^*}\mathfrak{M}_K\} \end{cases}$$

La consistance du facteur de Bayes sous H_1 est un problème bien étudié qui a été prouvé dans Ghosal et al. 2008 et Mcvinish et al. 2009 sous une hypothèse peu restrictive sur le modèle paramétrique et qui vaut pour les mélanges finis. Le résultat est obtenu en utilisant l'existence de tests exponentiellement consistants (Ghosh and Ramamoorthi 2003), ce qui fait que $BF_{0,1}$ converge exponentiellement rapidement vers 0 sous H_1.

Établir la consistance du facteur de Bayes sous H_0, cependant, est plus difficile. Un tel résultat a été obtenu dans Dass and Lee 2004 ou Verdinelli and Wasserman 1998 lorsqu'on considère une hypothèse nulle ponctuelle par rapport à une grande classe d'alternatives non paramétriques. Pour des hypothèses H_0 plus générales, Mcvinish et al. 2009 propose des conditions suffisantes sur les distributions non-paramétriques pour lesquelles la consistance du facteur de Bayes pour le test d'une famille paramétrique de distributions est vérifiée. Ces conditions nécessitent une compréhension pointue de la masse a priori des voisinages décroissants de f_0 sous la distribution a priori non-paramétrique impliquée par le processus de Dirichlet. D'autres approches consistent à considérer une distribution a priori non-paramétrique

modifiée sous H_1, comme dans Tokdar and Martin 2021 par exemple, où un processus de Dirichlet particulier est construit afin de tester la gaussianité d'un échantillon.

Outre les applications de type tests d'adéquation, la vraisemblance marginale du DPM peut être utilisée pour trouver la meilleure mesure mélangeante en minimisant le facteur de Bayes pour les alternatives paramétriques par rapport aux alternatives non paramétriques dans le contexte de l'estimation bayésienne de densité (Argiento et al. 2010) ou pour sélectionner des partitions appropriées de certaines données (Ray and Mallick 2006).

Contributions du chapitre 3

Supposons que, sous H_0, certaines données $\boldsymbol{y}$ proviennent d'un mélange fini P_{f_0} avec K_0 composantes. Nous notons sa densité par rapport à la mesure de Lebesgue par $f_0 := p(y|P_0) = \int f(y|\theta)dP_0(\theta)$, avec $P_0 = \sum_{k=1}^{K_0} \varpi_k^0 \delta_{\theta_k^0}$ et $\theta_k^0 \in \Theta \subset \mathbb{R}^d$ pour tout k.

Notre principale contribution est d'étudier le comportement asymptotique du facteur de Bayes

$$BF_{K,DP} = \frac{m_K(\boldsymbol{y})}{m_{DP}(\boldsymbol{y})}$$

où $m_K(\boldsymbol{y})$ est la vraisemblance marginale des données pour un modèle de mélange fini à K composantes, où K n'est pas nécessairement égal à K_0.
Notre première préoccupation est de prouver que le facteur de Bayes est consistant, c'est-à-dire

$$BF_{K_0,DP} \longrightarrow 0 \text{ en probabilité sous } P_{f_0}. \tag{1}$$

Établir la consistance du facteur de Bayes en comparant un modèle paramétrique comme les mélanges finis au modèle de mélange de processus de Dirichlet n'est pas immédiat. En particulier, il est nécessaire d'établir une borne supérieure asymptotique pour $m_{DP}(\boldsymbol{y})$. Nous nous concentrons sur la recherche d'une telle borne supérieure lorsque $f_0 \in \cup_{K\in\mathbb{N}^\star}\mathfrak{M}_K$, où $\mathfrak{M}_K$ désigne le modèle de mélanges avec K densités $f_\theta(y)$. Cela nécessite un contrôle précis de la masse a priori de voisinages L_1 décroissants autour de la vraie densité f_0. Nous fournissons un tel résultat dans le Lemme 0.1 ci-dessous.

Cela est fait sous un ensemble de cinq hypothèses, dont trois (**A1**–**A3**) concernent la régularité de la densité f et l'identifiabilité (forte) du mélange P_{f_0}. Une classe de modèles satisfaisant ces conditions est par exemple les mélanges de type position (*location*) de distributions Normales ou Laplace. L'hypothèse **A4** est une condition suffisante sur la mesure de base G_0 du processus de Dirichlet pour que les moyennes aléatoires $\mu(P) := \sum_{k=1}^{\infty} \varpi_k \theta_k$ existent presque sûrement, où $P \sim DP(M, G_0)$. Enfin, l'hypothèse **A5** exige que la distribution a priori Π_M sur M ait un support restreint à $[\zeta, \infty)$ pour certains $\zeta > 0$ et ait des queues exponentiellement décroissantes.

Lemma 0.1. *Soit $\boldsymbol{y} = (y_1, \ldots, y_n)$ des observations i.i.d de $f_0 := f_{P_0}$ avec $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$. Supposons que les hypothèses **A1**–**A3** sont satisfaites. Soit Π la distribution a priori conjointe sur (P, M) où $P|M \sim DP(M, G_0)$ et $M \sim \Pi_M$, de telle sorte que Π vérifie les hypothèses **A4**–**A5**. Alors, pour toute suite δ_n telle que $\delta_n \overset{n\to\infty}{\longrightarrow} 0$,*

$$\Pi\left(\|f_P - f_0\|_1 < \delta_n\right) \lesssim \delta_n^{K_0 - 1 + dK_0 + \zeta - \varepsilon}$$

où $\varepsilon > 0$ peut être choisi aussi proche de 0 que désiré.

La principale difficulté dans la preuve du lemme ci-dessus est d'établir que la densité des moyennes aléatoires de Dirichlet $\mu(P)$ a une densité continue et bornée, ce qui est fait en montrant que leur fonction caractéristique est intégrable. Le Lemme 0.1 implique ensuite le théorème suivant, qui fournit une borne supérieure asymptotique pour $m_{DP}(\boldsymbol{y})$.

Theorem 0.2. *Soit $\boldsymbol{y} = (y_1, \ldots, y_n)$ des observations i.i.d de $f_0 := f_{P_0}$ avec $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$. Supposons que les hypothèses **A1**–**A3** sont satisfaites. Soit Π la distribution a priori conjointe sur (P, M) où $P|M \sim DP(M, G_0)$ et $M \sim \Pi_M$, de telle sorte que Π vérifie les hypothèses **A4**–**A5**. Alors pour tout $0 < t < \zeta$,*

$$P_{f_0}\left(m_{DP}(\boldsymbol{y}) > n^{-(K_0 - 1 + dK_0 + t)/2}\right) \overset{n\to\infty}{\longrightarrow} 0$$

où

$$m_{DP}(\boldsymbol{y}) := \int f_P(\boldsymbol{y})/f_0(\boldsymbol{y}) d\Pi(P)$$

En utilisant les bornes inférieures calculées dans Rousseau and Mengersen 2011, le résultat ci-dessus se traduit directement en un résultat de consistance du facteur de Bayes, qui est donné dans le corollaire suivant.

Corollary 0.3. *Si la distribution a priori Π vérifie les hypothèses **A4**–**A5** et la distribution a priori sur $\mathfrak{M}_{K_0}$ est définie comme dans (2.5), alors*

*(i) si $f_{P_0} \in \mathfrak{M}_{K_0}$ satisfait les hypothèses **A1**–**A3**, alors $m_{K_0}(\boldsymbol{y})/m_{DP}(\boldsymbol{y}) \to \infty$ sous f_{P_0}.*

(ii) De plus, pour tout $K > K_0$, si la distribution a priori sur $\mathfrak{M}_K$ est définie comme dans (2.5) et a pour hyperparamètre de concentration $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$ pour un certain $\alpha > 0$, et que soit $\alpha < d/2 \wedge \zeta/(K - K_0)$, ou bien $d/2 < \alpha \wedge \zeta/(K - K_0)$, alors $m_K(\boldsymbol{y})/m_{DP}(\boldsymbol{y}) \to \infty$ sous f_{P_0}.

(iii) Si $\inf_{f_P \in \mathfrak{M}_K} KL(f_{P_0}, f_P) > 0$ et la distribution de Dirichlet a priori vérifie $\Pi_{DP}(KL(f_{P_0}, f_P) \leq \epsilon) > 0$ pour tout $\epsilon > 0$, alors $m_K(\boldsymbol{y})/m_{DP}(\boldsymbol{y}) \to 0$ sous f_{P_0}.

Notons que le résultat ci-dessus atteint plus que l'objectif d'établir la consistance du facteur de Bayes (1). En fait, les modèles de mélanges surestimés (c'est-à-dire lorsque $K > K_0$) sont favorisés par le facteur de Bayes sous un choix approprié des hyperparamètres sur les poids du mélange fini et une troncation suffisamment grande de la distribution a priori de M. Ce résultat supplémentaire est établi en utilisant les bornes inférieures asymptotiques fournies dans Rousseau and Mengersen 2011 sur $m_K(\boldsymbol{y})$ lorsque $K > K_0$.

En pratique, un tel test d'adéquation est intéressant si l'on est capable d'estimer correctement l'intégrale incalculable définissant $m_{DP}(\boldsymbol{y})$. À notre connaissance, seul les travaux de Basu and Chib 2003 abordent cette question en proposant une adaptation de l'algorithme de Chib de Chib 1995. Nous comparons cette méthode avec une nouvelle approche basée sur la régression logistique inverse (RLR) établie

dans Geyer 1994, ainsi qu'avec d'autres estimateurs. Nos résultats empiriques suggèrent que malgré ses bonnes performances, la méthode de Chib ne semble aussi robuste que la RLR lorsque la taille des données n augmente.

Dans l'ensemble, nous espérons que les résultats du Théorème 0.2 combinés à la nouvelle méthode d'estimation de $m_{DP}(\boldsymbol{y})$ aideront à éclairer le choix entre modèles paramétriques et non-paramétriques.

Calcul distribué de la vraisemblance marginale des les mélanges finis

Le calcul distribué a récemment émergé en tant que paradigme puissant pour résoudre des problèmes complexes en apprentissage automatique et en inférence statistique, lorsque la quantité de données est trop importante pour être traitée par une seule machine. Cette approche consiste à diviser la tâche d'inférence en sous-problèmes plus petits qui peuvent être résolus indépendamment par plusieurs machines ou processeurs, puis à combiner les résultats pour obtenir une solution globale au problème initial. Cela réduit non seulement la charge computationnelle sur les machines individuelles, mais facilite également le calcul parallèle, permettant une inférence plus rapide et plus efficace.

Du point de vue de l'inférence bayésienne, cette stratégie, généralement appelée *divide and conquer*, est principalement composée de trois étapes. Tout d'abord, les données $\boldsymbol{y}$ sont *divisées* en S lots non chevauchants $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S$ et la distribution a posteriori complète est décomposée de la façon suivante

$$\begin{aligned} \pi(\boldsymbol{\vartheta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) \\ &= \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} \\ &\propto \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) \end{aligned} \tag{2}$$

pour une fonction de vraisemblance $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ et une distribution a priori $\pi(\boldsymbol{\vartheta})$, où pour tout $s = 1, \ldots, S$, $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ est appelée la sous-distribution a posteriori du paramètre $\boldsymbol{\vartheta}$ sur le lot $\boldsymbol{y}_s$. L'inférence est ensuite réalisée sur chaque sous-ensemble de données indépendamment, éventuellement en parallèle, en exécutant des algorithmes MCMC sur plusieurs unités de calcul. Enfin, les différents échantillons des sous-distribution a posteriori sont *recombinés* pour approcher un échantillon de la distribution a posteriori *complète*. Cette dernière étape est généralement complexe car il n'existe pas de moyen exact de transformer une collection d'échantillons $\{\boldsymbol{\vartheta}_t^{(s)}\}_t$ provenant de chaque sous-distribution a posteriori $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ en un échantillon $\{\boldsymbol{\vartheta}_t\}_t$ distribué selon $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. En se basant sur le théorème de Bernstein-von Mises, Huang and Gelman 2005 et Scott et al. 2016 utilisent des approximations normales aux sous-distributions a posteriori pour reconstruire un échantillon de la distribution a posteriori complète comme une moyenne pondérée des échantillons de sous-distribution a posteriori. En supposant toujours la normalité des sous-distributions a posteriori, Neiswanger et al.

2014 utilise une estimation par noyau pour reconstruire la distribution a posteriori complète. D'autres approches consistent à recombiner les sous-échantillons via leur barycentre dans un espace de Wasserstein des mesures de probabilité (Srivastava et al. 2018), ou leur médiane géométrique (Minsker et al. 2014). Les applications de la stratégie *divide and conquer* sont nombreuses. En premier lieu, elle peut être utilisée pour réduire considérablement le coût de l'inférence a posteriori lorsque la quantité de données disponible est très importante, en réduisant les goulets d'étranglement en mémoire et en calcul. Une autre application immédiate est l'inférence sur une architecture de données distribuées, dans laquelle les données sont stockées à différents endroits, soit pour des raisons de confidentialité, par exemple avec des données de santé (Hallock et al. 2021), soit simplement parce que la taille de l'ensemble de données est trop grande pour être stockée sur une seule machine.

Malgré sa popularité, l'application du paradigme *divide and conquer* à la modélisation par mélanges reste largement inexplorée jusqu'à présent. En effet, bien que l'inférence sur chaque lot de données $\boldsymbol{y}_s$ ne pose généralement pas de problème, l'hypothèse habituelle de normalité asymptotique utilisée pour la recombinaison des S échantillons MCMC ne s'applique pas à la distribution a posteriori des modèles de mélange.

Dans le cas général, le défi du calcul distribué de la vraisemblance marginale d'un modèle selon une stratégie *divide and conquer* reste largement inexploré jusqu'à présent. Le problème réside dans le fait qu'il existe une manière pratique de relier la distribution a posteriori complète et les sous-distributions a posteriori par le biais de (2), mais qu'une telle identité ne s'applique pas à la vraisemblance marginale. En effet,

$$\begin{aligned} m(\boldsymbol{y}) &= \int \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta} \\ &\neq \prod_{s=1}^{S} \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}. \end{aligned}$$

Cela nécessite une approche différente pour combiner les estimations de vraisemblance marginale calculées sur chaque lot $\boldsymbol{y}_s$ séparément. Très récemment, une identité reliant $m(\boldsymbol{y})$ et $\{\tilde{m}(\boldsymbol{y}_s)\}_{s=1}^{S}$ a été formalisée dans Buchholz et al. 2022.

Proposition 0.4 (Buchholz et al. 2022)**.** *Soit $\boldsymbol{y}$ des données i.i.d et un modèle $\mathcal{P}$ pour lequel la fonction de vraisemblance se factorise en $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$, la vraisemblance marginale des données peut être écrite comme*

$$m(\boldsymbol{y}) = Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{3}$$

où pour chaque $s = 1, \ldots, S$,

$$\begin{aligned} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) &\propto p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta}), \\ \tilde{m}(\boldsymbol{y}_s) &= \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \end{aligned}$$

et

$$Z = \int \pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}$$

Pour la plupart des choix de prior $\pi(\boldsymbol{\vartheta})$, sous certaines contraintes potentielles sur les hyperparamètres, la distribution a priori $\tilde{\pi}(\boldsymbol{\vartheta})$ est une distribution de probabilité bien définie et l'intégrale Z est analytique. Dans le cas des mélanges finis, l'estimation des vraisemblances marginales sur chaque lot $\tilde{m}(\boldsymbol{y}_s)$ peut être réalisée en utilisant les algorithmes décrits au Chapitre 2. Cependant, l'intégrale $I := \int_\Theta \prod_{s=1}^S \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$ est clairement intractable pour les mélanges finis. Pour la classe plus générale des modèles *conditionnellement conjugués*, caractérisés par la disponibilité d'une expression en forme close de la distribution postérieure augmentée $\pi(\boldsymbol{\vartheta}|\boldsymbol{z}, \boldsymbol{y})$ pour des variables latentes $\boldsymbol{z}$, Buchholz et al. 2022 proposent un estimateur prometteur pour I.

Proposition 0.5 (Buchholz et al. 2022)**.** *Pour des modèles conditionnellement conjugués, l'intégrale $I := \int_\Theta \prod_{s=1}^S \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$ peut être estimée par*

$$\hat{I} = \frac{1}{T} \sum_{t=1}^{T} \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s^{(t)}, \boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4}$$

où $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ et $\hat{I}$ est un estimateur non biaisé de I.

Malgré la nature conditionnellement conjuguée des mélanges finis, nous argumentons que l'estimateur (4) ne convient pas pour ces modèles. Nous proposons des versions "corrigées" de $\hat{I}$ ainsi qu'une stratégie SMC (Sequential Monte Carlo) pour résoudre le problème d'estimation distribuée de la vraisemblance marginale des modèles de mélange finis.

Contributions du Chapitre 4

Nous remarquons que l'utilisation de la variable latente d'appartenance au groupe $\boldsymbol{z}$ dans l'astuce d'augmentation des données suggérée par l'équation (4) rend l'estimation de I à travers $\hat{I}$ difficile dans le cas des modèles de mélanges finis. Le problème ici est que les étiquettes de groupe $\boldsymbol{z}_s$ ne sont pas cohérentes entre les S lots. Bien que cela ne rende pas l'estimateur $\hat{I}$ biaisé, cela augmente certainement sa variance dans le cas où l'échantillonneur de Gibbs ne visite pas uniformément toutes les $K!$ configurations modales, au sein *et* entre les S lots, ce qui ne se produit que rarement étant donné un budget computationnel fini. Nous proposons une solution simple, dans la même veine que la correction proposée par Berkhof et al. 2003 pour l'estimateur de Chib (Chib 1995), qui consiste à faire la moyenne sur toutes les permutations des variables latentes $\boldsymbol{z}_s$, pour tout $s = 1, \ldots, S$. Ceci est donné dans la proposition suivante.

Proposition 0.6 (Estimateur permuté de I)**.** *Soit S un entier supérieur à 1 et*

$\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ *pour* $s = 1, \ldots, S$. *Alors*

$$\hat{I}_{perm} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2,\ldots,\sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_1^{(t)}, \boldsymbol{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s^{(t)}), \boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{5}$$

est un estimateur non biaisé de I.

Malgré le fait que l'estimateur $\hat{I}_{perm}$ compense la variance de Monte Carlo explosive de $\hat{I}$, il a un coût computationnel élevé de l'ordre de $O\big(TK!^{S-1}\big)$. Afin de réduire ce dernier, nous proposons une stratégie d'échantillonnage préférentiel qui sélectionne les combinaisons de permutations $(\sigma_2, \ldots, \sigma_S)$ qui contribuent le plus à l'intégrale de l'équation (5). Cela réduit considérablement le coût de $\hat{I}_{perm}$ à $O(T(Kn/S + K! + M))$, où M est le nombre de simulations d'importance à chaque itération $t = 1, \ldots, T$. Malgré cet estimateur corrigé et relativement peu coûteux pour I, nous nous rendons compte que la variance intrinsèque due à l'astuce d'augmentation (4) est trop grande dans la plupart des situations, en raison de l'écart entre les distributions $\tilde{\pi}(\boldsymbol{z}|\boldsymbol{y}) := \prod_{s=1}^S \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$, à partir desquelles les variables latentes sont échantillonnées, et la "véritable" distribution postérieure complète $\pi(\boldsymbol{z}|\boldsymbol{y})$.

Pour cette raison, nous changeons notre approche du problème en considérant simplement l'intégrale I comme la constante de normalisation d'un produit non normalisé de distributions. En effet, en définissant le produit incrémental de distributions

$$\tilde{\pi}_s(\boldsymbol{\vartheta}) \propto \prod_{l=1}^{s} \tilde{\pi}(\boldsymbol{\vartheta}_l|\boldsymbol{y}_l)$$

nous remarquons que I est simplement égal à $Z_S = \int_\Theta \tilde{\pi}_S(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$. Cela permet de réécrire l'identité originale (3) dans Buchholz et al. 2022 comme dans la proposition suivante.

Proposition 0.7 (Une nouvelle identité)**.** *Soit* $\boldsymbol{y}$ *des données i.i.d et un modèle statistique pour lequel la fonction de vraisemblance se factorise en* $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^S p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$*, la vraisemblance marginale des données peut être décomposée comme*

$$m(\boldsymbol{y}) = Z^S \times \tilde{m}(\boldsymbol{y}_1) \times \prod_{s=2}^{S} \int \pi_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$$

Maintenant, en remarquant que pour tout $s = 2, \ldots, S$,

$$\int \tilde{\pi}_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \mathbb{E}_{\tilde{\pi}_{s-1}}(p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}))$$

découle une stratégie SMC pour estimer $m(\boldsymbol{y})$ en utilisant la séquence de distributions $\{\tilde{\pi}_{s-1}(\boldsymbol{\vartheta})\}_{s=2}^S$. Non seulement cette méthode montre une performance améliorée empiriquement, en termes de variance et de temps de calcul, mais elle s'applique également à une classe de modèles plus générale, en relâchant toute hypothèse de conjugalité. Nous sommes optimistes quant au fait que cette approche pourrait faciliter la tâche d'estimation de la vraisemblance marginale de grands ensembles de

données et/ou distribués au sein de plusieurs entités, même en dehors du contexte de la modélisation par mélanges.

Chapter 1

Introduction

1.1 Evidence estimation for Finite mixtures

Mixture models are of significant interest due to their convenient way of modeling heterogeneity in a given population. As such, they were first formally introduced by Pearson 1894 in which a two-component Normal mixture is fitted to the ratio of forehead to body-length in a population of crabs. By matching moments and solving a polynomial equation of degree nine, Pearson identifies two homogeneous clusters within the population of crabs, shedding light on a probable evolutionary divergence between the two groups, as shown on Figure 1.1.

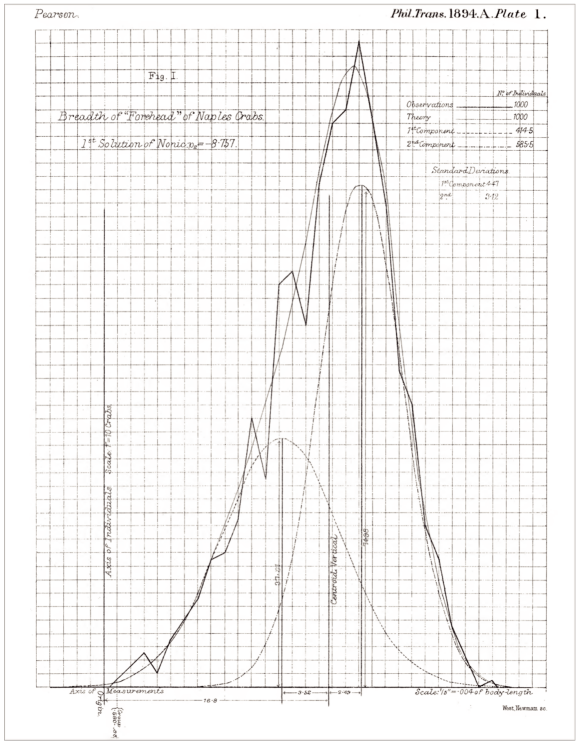


Figure 1.1: The two component mixture as fitted by Pearson 1894 on the population of crabs. The solid line is a *frequency curve*. The dashed lines represent the two weighted normal mixture components.

1.1.1 Model specification and model order selection

Formally, we say that some independent and identically distributed (*i.i.d*) data $\boldsymbol{y} = (y_1, \ldots, y_n)$ arise from a Finite Mixture model (FM) with K components if their density can be written as

$$p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \varpi_k f(y_i|\theta_k)$$

for $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi}) \in \Theta^K \times \Delta_{K-1}$ where $\Theta \subset \mathbb{R}^d$ and we denote by $\Delta_{K-1} := \{\boldsymbol{\varpi} \in (0,1)^K : \sum_{k=1}^{K} \varpi_k = 1\}$ the $(K-1)$-dimensional simplex. The vector $(\varpi_1, \ldots, \varpi_K)$ contains the so-called *mixture weights* while the *mixture parameters* $(\theta_1, \ldots, \theta_K)$ parametrize some density f that is usually called *mixture kernel*.

Mixture modeling can be for the most part reduced to choosing an appropriate mixture kernel f and the right model order K, which corresponds to the number of mixture components. The latter typically requires careful selection and/or estimation and is of particular interest when the main goal of inference is related to clustering. Mixture models have indeed become one of the main tools used by practitioners who wish to identify sub-groups with specific characteristics in a population (McLachlan and Basford 1988; Geary 1989; Dang et al. 2023). Similar but more involved applications include topic modeling such as Latent Dirichlet allocation (Blei et al.

2003; Chen and Doss 2019), or image segmentation, compression and classification (Aiyer et al. 2005; Zeng et al. 2014).

Existing methods for estimating K can be found in the thorough review by Celeux et al. 2019. They include frequentist strategies such as suitably adapted Likelihood Ratio Tests (LRTs, McLachlan 1987; Heckman et al. 1990; McLachlan and Peel 2000; Frühwirth-Schnatter 2006), method of moments estimators (Dacunha-Castelle and Gassiat 1997), or popular information criteriae (Smyth 2000).

From a Bayesian perspective, estimating K can be done alongside the estimation of the mixture parameters and weights $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi})$ by setting a prior distribution on K. Sampling from the subsequent posterior distribution of K is far from straightforward since it requires to design so-called *transdimensional* samplers. The seminal work by Green 1995 and Richardson and Green 1997 led to the construction of Reversible Jump Monte Carlo Markov Chains (RJMCMC) which allow such 'jumps' between spaces of different dimensions. Such models where K is treated as any other parameter have been formalized by Richardson and Green 1997 or Nobile 2004 where they were referred to as mixtures with a random number of components. The term Mixture of Finite Mixtures (MFM, Miller and Harrison 2018) is now more generally used by practitioners. Recent advances include the derivation of new sampling strategies for such complex models (see for instance Frühwirth-Schnatter et al. 2021).

Another Bayesian approach consists in computing the Bayes Factor (Jeffreys 1935; Raftery 1996) for competing values of K. Such a strategy is known to be consistent (i.e the Bayes Factor consistently points towards the right model for an increasing number of observations n, Chib and Kuffner 2016). In practice, concrete applications of this approach require the estimation of the marginal likelihood of a finite mixture model, defined as the integral of the likelihood function against the prior distribution, which is not an easy task. The most popular Monte Carlo algorithms to tackle this challenge are Chib's algorithm (Chib 1995), Bridge Sampling (Meng and Wong 1996; Frühwirth-Schnatter 2004; Frühwirth-Schnatter 2019), and to a lesser extent, Sequential Monte Carlo methods (Chopin 2002; Gunawan et al. 2020).

More generally, being able to compute the marginal likelihood of a model is of crucial importance in a Bayesian setting as it is the main tool used for model evaluation and comparison. In fact, Fong and Holmes 2020 shows that computing the marginal likelihood of a model is formally equivalent to exhaustive leave-p-out cross-validation averaged over all values of p, cross-validation being the gold standard frequentist procedure for model evaluation. Hence, being able to compute the marginal likelihood of a Finite mixture model is of interest not only for consistently selecting a value of K, but also for finding a suitable mixture kernel f, or even comparing a mixture model against a different parametric or nonparametric alternative.

1.1.2 Contributions of Chapter 2

Successfully approximating the marginal likelihood of mixture models requires *ad hoc* methods which account for the specific challenges that arise when dealing with such models. In particular, for an exchangeable prior on the mixture weights and

parameters, the mixture posterior distribution $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$ is permutation invariant. This translates to

$$\pi((\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K)|\boldsymbol{y}) = \pi((\boldsymbol{\vartheta}_{\sigma(1)}, \ldots, \boldsymbol{\vartheta}_{\sigma(K)}|\boldsymbol{y})$$

for all $(\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K) \in \Theta^K \times \Delta_{K-1}$ and all $\sigma \in \mathfrak{S}_k$, where $\mathfrak{S}_K$ denotes the set of permutations of $\{1, \ldots, K\}$.
This implies that the posterior distribution displays $K!$ equiprobable modes. As noticed by Neal 1999 in a response to Chib 1995, a good mixing behavior of MCMC algorithms targeting the posterior distribution is essential for meaningful approximation of the model evidence. However, in the case of finite mixtures, this requires a balanced visit of all $K!$ modal configurations of the posterior (a phenomenon also called *label switching*), which, most of the time, is an unrealistic expectation given a finite computational budget. To make up for this ill-behavior of MCMC samplers, popular methods such as Chib's algorithm and Bridge Sampling and their subsequent adaptations to finite mixtures resort to artificially enforcing a perfect label switching phenomenon by integrating their estimators over the space of permutations $\mathfrak{S}_K$. This comes at an additional cost of $O(K!)$ which is barely sustainable even for values of K as small as 5. In Chapter 2, we first introduce a modified Chib's algorithm which makes use of the partitioning structure induced by mixture models. This quantity is of significant interest since it is resilient to label switching, thus avoiding the exponential cost of $O(K!)$ paid by traditional methods. We also adapt the sequential imputation algorithm discovered by Kong et al. 1994 to finite mixtures. On top of being robust to label switching, this approach also proves to scale very well with the number of observations n. Subsequently, we provide an empirical review of classical estimators of the marginal likelihood, old and new, and highlight their strengths and weaknesses in different scenarios where K and n need not be small. In particular, we find that Sequential Monte Carlo approaches are the most effective and we are hopeful this assessment will incite practitioners to use these methods more often.

1.2 Evidence asymptotics and estimation for the Dirichlet Process mixture model

1.2.1 Model spectification

The Dirichlet Process Mixture model (DPM), first introduced by Ferguson 1983 is one of the main tools of the field of Bayesian nonparametrics. Assuming an infinite number of mixture components a priori, it is sometimes called the 'infinite' mixture model as opposed to finite mixtures. Indeed, for some $i.i.d$ data $\boldsymbol{y} = (y_1, \ldots, y_n)$, the DPM can be described by the following generative model specification

$$\begin{aligned} y_i|\theta_i &\overset{i.i.d}{\sim} F(y_i|\theta_i) \quad \text{for } i = 1, \ldots, n \\ \theta_i|P &\sim P \\ P|M &\sim DP(M, G_0) \\ M &\sim \Pi_M \end{aligned}$$

where $F(\cdot|\theta)$ is a distribution supported on $\mathbb{R}^d$ with density with respect to the Lebesgue measure $f(\cdot|\theta)$ and $DP(M, G_0)$ denotes the Dirichlet Process with base measure G_0 and concentration parameter M. The realizations P of the Dirichlet Process are almost surely discrete such that, if δ_x denotes the Dirac delta function,

$$P = \sum_{k=1}^{\infty} \varpi_k \delta_{\theta_k}$$

with $\varpi_k = V_k \prod_{l<k}(1 - V_l)$ for some $V_l \overset{i.i.d}{\sim} Beta(1, M)$ (ensuring that $\sum_{k=1}^{\infty} \varpi_k = 1$ almost surely) and $\theta_1, \theta_2, \dots \overset{i.i.d}{\sim} G_0$.
The likelihood of P can then be written as

$$f_P(\boldsymbol{y}) := p(\boldsymbol{y}|P) = \prod_{i=1}^{n} \int_{\Theta} f(y_i|\theta) dP(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{\infty} \varpi_k f(y_i|\theta_k).$$

The range of applications of the DPM is broad, from multivariate clustering (Crépet and Tressou 2011) to Bayesian density estimation (Neal 1992; Rabaoui et al. 2012). An appealing trait of the DPM is the very mild prior assumption set on the distribution of the mixture weights $\boldsymbol{\varpi} = (\varpi_1, \varpi_2, \dots)$ and parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots)$ since their prior P is itself random, making it one of the cornerstones of Bayesian nonparametrics.

1.2.2 Bayes factor consistency for testing a parametric null hypothesis against a non-parametric Dirichlet Process mixture alternative

Resorting to this kind of flexible, nonparametric models may be motivated by a wish to free oneself from the constraint of model selection and comparison. Therefore, it might seem surprising at first to care about the estimation of the model evidence of the DPM, which can be defined as

$$m_{DP}(\boldsymbol{y}) = \int f_P(\boldsymbol{y}) \Pi(dP, dM)$$

where Π denotes the joint prior distribution on (P, M) with $P|M \sim DP(M, G_0)$ and $M \sim \Pi_M$.
However, it happens that the quantity $m_{DP}(\boldsymbol{y})$ can have several applications of interest. For instance, the usual problem of statistical modeling of determining whether an $n-$sample $\boldsymbol{y} = (y_1, \dots, y_n)$ arises from a particular parametric family of distributions can be formalized as a goodness of fit test against a nonparametric alternative, like the DPM. Formally, we consider here the problem of testing the parametric null hypothesis $H_0 : f_0 \in \cup_{K \in \mathbb{N}^*} \mathfrak{M}_K$, where $\mathfrak{M}_K$ denotes the parametric model of mixtures with K emission distributions $f_\theta(y)$, against the alternative $H_1 : f_0 \in \mathcal{F} \setminus \{\cup_{K \in \mathbb{N}^*} \mathfrak{M}_K\}$ modelized by the Dirichlet Process mixture model with mixture kernel $f_\theta(y)$, for $\mathcal{F}$ an encompassing set of density functions. The latter procedure is relevant from a Bayesian perspective provided that the Bayes Factor is

consistent, that is

$$\lim_{n\to\infty} BF_{0,1} := \lim_{n\to\infty} \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} = \begin{cases} \infty & \text{under } H_0 : f_0 \in \cup_{K\in\mathbb{N}^*}\mathfrak{M}_K \\ 0 & \text{under } H_1 : f_0 \in \mathcal{F} \setminus \{\cup_{K\in\mathbb{N}^*}\mathfrak{M}_K\} \end{cases}$$

The consistency of the Bayes Factor under H_1 is a well-studied problem that was proved by Ghosal et al. 2008 and Mcvinish et al. 2009 under a mild assumption on the parametric model that holds for finite mixtures. The result is obtained using the existence of exponentially consistent tests (Ghosh and Ramamoorthi 2003), which in turn makes $BF_{0,1}$ converge to 0 exponentially fast under H_1.

Establishing the consistency of the Bayes Factor under H_0, however, is more challenging. Such a result was obtained by Dass and Lee 2004 or Verdinelli and Wasserman 1998 when considering a point null hypothesis against a large class of nonparametric alternatives. For more general hypotheses H_0, Mcvinish et al. 2009 derive sufficient conditions on nonparametric distributions for which the consistency of the Bayes factor for testing a parametric family of distributions holds. These conditions require a refined understanding of the prior mass of decreasing neighborhoods of f_0 under the Dirichlet Process non-parametric prior. Other approaches consist in considering a modified nonparametric prior under H_1 as in Tokdar and Martin 2021 for example, where a particular DPM alternative is constructed for testing normality.

Besides goodness-of-fit type of applications, the marginal likelihood of the DPM can be used to find the best-fitting mixing measure by minimizing the Bayes Factor for parametric against nonparametric alternatives in the context of Bayesian density estimation (Argiento et al. 2010) or to select suitable partitions of some data (Ray and Mallick 2006).

1.2.3 Contributions of Chapter 3

Let us assume that, under H_0, some data $\boldsymbol{y}$ arise from a finite mixture P_{f_0} with K_0 components. We denote its density with respect to the Lebesgue measure by $f_0 := p(y|P_0) = \int f(y|\theta)dP_0(\theta)$, with $P_0 = \sum_{k=1}^{K_0} \varpi_k^0 \delta_{\theta_k^0}$ and $\theta_k^0 \in \Theta \subset \mathbb{R}^d$ for all k.

Our main contribution is to study the asymptotic behavior of the Bayes Factor

$$BF_{K,DP} = \frac{m_K(\boldsymbol{y})}{m_{DP}(\boldsymbol{y})}$$

where $m_K(\boldsymbol{y})$ is the marginal likelihood of the data for a $K-$component finite mixture model, where K need not be equal to K_0.
Our first concern is to prove that the Bayes Factor is consistent, that is

$$BF_{K_0,DP} \longrightarrow 0 \text{ in } P_{f_0} \text{ probability.} \tag{1.1}$$

Establishing the consistency of the Bayes Factor comparing a parametric model like finite mixtures against the Dirichlet Process mixture model is not straightforward. In particular, one needs to derive an asymptotic upper-bound on $m_{DP}(\boldsymbol{y})$. We concentrate on deriving such an upper bound when $f_0 \in \cup_{K\in\mathbb{N}^\star}\mathfrak{M}_K$, where $\mathfrak{M}_K$

denotes the model of mixtures with K densities $f_\theta(y)$. This requires a refined control of the a priori mass of decreasing L_1-neighborhoods around the true density f_0. We provide such a result in Lemma 3.3 below.

This is done under a set of five assumptions, three of which (**A1**–**A3**) concern the regularity of the kernel density f and the (strong) identifiability of the mixture P_{f_0}. A class of models satisfying these conditions are for instance location mixtures of Normal or Laplace distributions. Assumption **A4** is a sufficient condition on the Dirichlet Process prior base measure G_0 for Dirichlet random means $\mu(P) := \sum_{k=1}^{\infty} \varpi_k \theta_k$ to exist almost surely, where $P \sim DP(M, G_0)$. Finally, assumption **A5** requires that the prior Π_M on M has support on $[\zeta, \infty)$ for some $\zeta > 0$ and has exponentially decreasing tails.

Lemma 3.3. *Assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ are i.i.d observations from $f_0 := f_{P_0}$ with $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$ and that Assumptions **A1**–**A3** are satisfied. Denote by Π the joint prior distribution on (P, M) where $P|M \sim DP(M, G_0)$ and $M \sim \Pi_M$, such that Π verifies Assumptions **A4**–**A5**. Then for any sequence δ_n such that $\delta_n \overset{n\to\infty}{\longrightarrow} 0$,*

$$\Pi\left(\|f_P - f_0\|_1 < \delta_n\right) \lesssim \delta_n^{K_0 - 1 + dK_0 + \zeta - \varepsilon}$$

where $\varepsilon > 0$ can be chosen as close to 0 as desired.

The main difficulty in the proof of the above lemma is to establish that the density of the Dirichlet random means $\mu(P)$ has a continuous and bounded density, which is done by showing that their characteristic function is integrable. Lemma 3.3 subsequently implies the following Theorem, which provides an asymptotic upper-bound on $m_{DP}(\boldsymbol{y})$.

Theorem 3.1. *Assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ are i.i.d observations from $f_0 := f_{P_0}$ with $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$ and that Assumptions **A1**–**A3** are satisfied. Denote by Π the joint prior distribution on (P, M) where $P|M \sim DP(M, G_0)$ and $M \sim \Pi_M$, such that Π verifies Assumptions **A4**–**A5**. Then for any $0 < t < \zeta$,*

$$P_{f_0}\left(m_{DP}(\boldsymbol{y}) > n^{-(K_0 - 1 + dK_0 + t)/2}\right) \overset{n\to\infty}{\longrightarrow} 0$$

where

$$m_{DP}(\boldsymbol{y}) := \int f_P(\boldsymbol{y}) / f_0(\boldsymbol{y}) d\Pi(P)$$

The above result directly translates into a Bayes Factor consistency result, which is given in the following corollary

Corollary 3.2. *If the DP prior verifies **A4**–**A5** and the prior on $\mathfrak{M}_{K_0}$ is defined as in (2.5), then*

*(i) If $f_{P_0} \in \mathfrak{M}_{K_0}$ satisfies Assumptions **A1**–**A3**, then $m_{K_0}(\mathbf{y})/m_{DP}(\mathbf{y}) \to \infty$ under f_{P_0}.*

(ii) Moreover for all $K > K_0$, if the prior on $\mathfrak{M}_K$ is defined as in (2.5) with Dirichlet hyperparameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$ for some $\alpha > 0$, if either $\alpha < d/2 \wedge \zeta/(K - K_0)$, or $d/2 < \alpha \wedge \zeta/(K - K_0)$, then $m_K(\mathbf{y})/m_{DP}(\mathbf{y}) \to \infty$ under f_{P_0}.

(iii) If $\inf_{f_P \in \mathfrak{M}_K} KL(f_{P_0}, f_P) > 0$ *and the DP prior verifies* $\Pi_{DP}(KL(f_{P_0}, f_P) \leq \epsilon) > 0$ *for all* $\epsilon > 0$*, then* $m_K(\mathbf{y})/m_{DP}(\mathbf{y}) \to 0$ *under* f_{P_0}*.*

Note that the result above achieves more than the goal of establishing the Bayes Factor consistency (1.1). In fact, overfitted mixture models (i.e when $K > K_0$) are favored by the Bayes Factor under a suitable choice of hyperparameters on the weights of the finite mixture and a large enough truncation of the prior on M. This additional result is established using the asymptotic lower-bounds provided in Rousseau and Mengersen 2011 on $m_K(\boldsymbol{y})$ when $K > K_0$.

In practice, such a goodness-of-fit test is of interest if one is able to correctly estimate the intractable integral defining $m_{DP}(\boldsymbol{y})$. To our knowledge, only Basu and Chib 2003 address this issue by proposing an adaptation of Chib's algorithm from Chib 1995. We compare this method with a new approach based on the Reverse Logistic Regression (RLR) trick established by Geyer 1994, as well as other estimators. We empirically find that despite its good performance, Chib's method does not seem to scale as well as RLR with an increase in the data size n.

Overall, we are hopeful that the results of Theorem 3.1 combined with the new estimation method of $m_{DP}(\boldsymbol{y})$ we suggest will ease the task of Bayesian model selection for practitioners. Our empirical results suggest that a result similar to Corollary 3.2 could be obtained for more general models, such as location-scale mixtures. The difficulty is to control the a priori mass of decreasing neighborhoods of f_0 in this scenario (as in Lemma 3.3) since it requires a refined understanding of linear functionals of the Dirichlet Process.

1.3 Distributed evidence computation for finite mixtures

Distributed computation has emerged as a powerful paradigm for tackling complex problems in machine learning and statistical inference, where the amount of data is too large to be processed by a single machine. This approach involves breaking down the inference task into smaller sub-problems that can be solved independently by multiple machines or processors, and then combining the results to obtain a global solution to the initial problem. This not only reduces the computational burden on individual machines but also facilitates parallel computation, enabling faster and more efficient inference.

From a Bayesian inference perspective, this strategy, usually called *divide and conquer*, is mainly composed of three steps. First, the data $\boldsymbol{y}$ is *divided* into S non-overlapping batches $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S$ and the full posterior distribution is decomposed as

$$\begin{aligned} \pi(\boldsymbol{\vartheta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) \\ &= \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} \\ &\propto \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) \end{aligned} \tag{1.2}$$

for a likelihood function $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ and a prior $\pi(\boldsymbol{\vartheta})$, where for all $s = 1, \ldots, S$, $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ is called the sub-posterior distribution of parameter $\boldsymbol{\vartheta}$ on batch $\boldsymbol{y}_s$. Posterior inference is then conducted on each subset of the data independently, possibly in parallel, running MCMC algorithms across several computing units. Lastly, the different sub-posterior samples are *recombined* to approximate a sample from the *full* posterior distribution. This last step is usually complex as there exists no exact way to transform a collection of samples $\{\boldsymbol{\vartheta}_t^{(s)}\}_t$ from each sub posterior $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ into a sample $\{\boldsymbol{\vartheta}_t\}_t$ that is distributed according to $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. Based on the Bernstein-von Mises theorem, Huang and Gelman 2005 and Scott et al. 2016 use normal approximations to the sub-posterior distributions to reconstruct a sample from the full posterior as a weighted average of the sub-posterior samples. Still assuming normality of the sub-posterior distributions, Neiswanger et al. 2014 uses kernel density estimation to reconstruct the full posterior. Other approaches consist in recombining the sub-samples through their barycenter in a Wasserstein space of probability measures (Srivastava et al. 2018), or their geometric median (Minsker et al. 2014). Applications of a *divide and conquer* strategy are numerous. First and foremost, it can be used in order to significantly reduce the cost of posterior inference when the amount of data at hand is very large by reducing memory and computational bottle necks. Another immediate application is inference on a distributed data architecture, in which data is stored at different locations, either for privacy issues, for instance with health data (Hallock et al. 2021), or simply because the data set size is too large to be stored on a single machine.

Despite its popularity, the application of the *divide and conquer* paradigm to mixture modeling remains largely unexplored so far. Indeed, while conducting inference on each data batch $\boldsymbol{y}_s$ is generally not an issue, the usual assumption of asymptotic normality used for recombination of the S MCMC samples does not hold for the posterior distribution of mixture models.

In the general case, the challenge of distributed computation of the marginal likelihood of a model in a *divide and conquer* fashion remains largely unexplored so far. The issue being that while there exists a convenient way of linking the full posterior distribution and the sub-posterior distribution through (1.2), such an identity does not hold for the marginal likelihood. Indeed,

$$\begin{aligned} m(\boldsymbol{y}) &= \int \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta} \\ &\neq \prod_{s=1}^{S} \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}. \end{aligned}$$

This calls for a different approach to combining marginal likelihoods estimates computed on each batch $\boldsymbol{y}_s$ separately. Very recently, an identity bridging the gap between $m(\boldsymbol{y})$ and $\{\tilde{m}(\boldsymbol{y}_s)\}_{s=1}^S$ has been derived by Buchholz et al. 2022.

Proposition 4.1 (Buchholz et al. 2022)**.** *For some data $\boldsymbol{y}$ and a model $\mathcal{P}$ for which the likelihood function factorizes as $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^S p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$, the marginal likelihood*

of the data can be written as

$$m(\boldsymbol{y}) = Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4.2}$$

where for each $s = 1, \ldots, S$,

$$\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) \propto p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta}),$$
$$\tilde{m}(\boldsymbol{y}_s) = \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta},$$

and

$$Z = \int \pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}$$

For most prior choices $\pi(\boldsymbol{\vartheta})$, under some potential constraints on the hyperparameters, the prior $\tilde{\pi}(\boldsymbol{\vartheta})$ is a well-defined probability distribution and the integral Z is analytical. In the case of finite mixtures, the estimation of the sub-marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ can be done using the algorithms described in Chapter 2. Integral $I := \int_\Theta \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$ however, is clearly intractable for finite mixtures. For the more general class of *conditionally conjugate* models, which are characterized by the availability of a closed-form expression for the augmented posterior $\pi(\boldsymbol{\vartheta}|\boldsymbol{z}, \boldsymbol{y})$ for some latent variable $\boldsymbol{z}$, Buchholz et al. 2022 derive a promising estimator for I.

Proposition 4.5 (Buchholz et al. 2022)**.** *For general conditionally conjugate models, the integral* $I := \int_\Theta \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$ *can be estimated through*

$$\hat{I} = \frac{1}{T} \sum_{t=1}^{T} \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s^{(t)}, \boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4.3}$$

where $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^{T} \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ *and* $\hat{I}$ *is an unbiased estimator of* I.

Despite the conditionally conjugate nature of finite mixtures, we argue that estimator (4.3) is not suited for these models. We propose 'corrected' versions of $\hat{I}$ as well as a SMC strategy to tackle the issue of distributed evidence computation for finite mixtures.

1.3.1 Contributions of Chapter 4

We remark that using the latent cluster membership variable $\boldsymbol{z}$ in the data augmentation trick suggested by equation (4.3) makes the estimation of I through $\hat{I}$ difficult in the case of finite mixture models. The issue here is that cluster labels $\boldsymbol{z}_s$ are not coherent across the S batches. While this does not make estimator $\hat{I}$ biased, it certainly makes its variance explode in the case where the Gibbs sampler does not visit evenly all $K!$ modal configurations, within *and* across the S batches, which clearly hardly ever happens given a finite computational budget. We propose an easy fix, in the same vein as the correction proposed by Berkhof et al. 2003 for Chib's

estimator (Chib 1995), which consists in averaging over all permutations of the latent variables $\boldsymbol{z}_s$, for all $s = 1, \dots, S$. This is given in the following proposition.

Proposition 4.6 (Permuted estimator of I)**.** *Let S be an integer larger than 1 and let $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ for $s = 1, \dots, S$. Then*

$$\hat{I}_{perm} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2,\dots,\sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_1^{(t)}, \boldsymbol{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s^{(t)}), \boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4.5}$$

is an unbiased estimator for I.

Despite making up for the explosive Monte Carlo variance of $\hat{I}$, estimator $\hat{I}_{perm}$ comes at the heavy computational cost of $O\big(TK!^{S-1}\big)$. In order to reduce the latter, we come up with an Importance Sampling strategy that selects the combinations of permutations $(\sigma_2, \dots, \sigma_S)$ that contribute the most to the integral of equation (4.5). This greatly reduces the cost of $\hat{I}_{perm}$ to $O(T(Kn/S + K! + M))$ where M is the number of importance simulations at each iteration $t = 1, \dots, T$.
Despite having derived a corrected and relatively cheap estimator for I, we realize that the intrinsic variance due to the augmentation trick (4.3) is too large in most situations, due to the discrepancy between the distributions $\tilde{\pi}(\boldsymbol{z}|\boldsymbol{y}) := \prod_{s=1}^S \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$, from which the latent variables are sampled, and the full 'true' posterior distribution $\pi(\boldsymbol{z}|\boldsymbol{y})$.

For this reason, we change our approach to the problem by simply regarding integral I as the normalizing constant of an unnormalized product of distributions. Indeed, by defining the incremental product of distributions

$$\tilde{\pi}_s(\boldsymbol{\vartheta}) \propto \prod_{l=1}^{s} \tilde{\pi}(\boldsymbol{\vartheta}_l|\boldsymbol{y}_l)$$

we remark that I is simply equal to $Z_S = \int_\Theta \tilde{\pi}_S(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$. This enables to rewrite the original identity (4.2) by Buchholz et al. 2022 as in the following proposition.

Proposition 4.8 (A new identity)**.** *For some data $\boldsymbol{y}$ and a statistical model for which the likelihood function factorizes as $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^S p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$, the marginal likelihood of the data can be decomposed as*

$$m(\boldsymbol{y}) = Z^S \times \tilde{m}(\boldsymbol{y}_1) \times \prod_{s=2}^{S} \int \pi_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$$

Now by remarking that for all $s = 2, \dots, S$,

$$\int \tilde{\pi}_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \mathbb{E}_{\tilde{\pi}_{s-1}}(p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}))$$

one can derive a SMC strategy to estimate $m(\boldsymbol{y})$ using the sequence of distributions $\{\tilde{\pi}_{s-1}(\boldsymbol{\vartheta})\}_{s=2}^S$. Not only does this method show enhanced performance empirically, both in terms of variance and computational time, but it also is applicable to a more general class of models, by relaxing any conjugacy assumption. We are hopeful this

approach could ease the estimation task of the marginal likelihood of large and/or distributed datasets, even outside of the context of mixture modeling.

Chapter 2

Evidence estimation for finite mixtures

Abstract

In this chapter, we consider the problem of estimating the marginal likelihood of finite mixture models by Monte Carlo methods. We review classical methods and highlight their expensive computational cost as soon as the number of mixture components K is bigger than 5. We then propose alternative algorithms that show better scaling properties.

Contents

2.1 Introduction

Mixture models are of significant interest due to their convenient way of modeling heterogeneity in a given population. Essentially, when some group structure is present in the data, Finite Mixture models (FM) arise as a natural tool to conduct a Bayesian clustering analysis, using for instance a collapsed Gibbs sampler (Algorithm 2). They have been successfully applied to document classification (Blei et al. 2003) but also more generally to computer vision, genetics, physics, or economics (see McLachlan et al. 2019 for a comprehensive review of applications).
Most difficulties in applying Finite Mixtures to real-world data reduce to the choice of mixture kernels and/or the number of components K. Celeux et al. 2019 lists the different existing strategies to estimate K. From a frequentist perspective, the most natural approach is to resort to likelihood ratio tests (LRT). As highlighted by McLachlan and Peel 2000, this procedure is not immediately applicable to finite mixtures due to identifiability and regularity issues. Indeed, when testing say $H_0 = K$ vs $H_1 = K + 1$, the null hypothesis is specified by parameters lying on the boundary of the parameter space, making standard results about LRTs invalid. Bootstrap strategies (McLachlan 1987) or direct modifications of the LRT (Frühwirth-Schnatter 2006) have been applied to circumvent this challenge, to name a few. Other approaches involve method of moments estimators (Heckman et al. 1990; Dacunha-Castelle and Gassiat 1997) or information criteria (Smyth 2000), the latter being popular alternatives to LRT, which, in real-life applications, are not commonly used by practitioners due to the complexity of their implementation for Finite Mixtures.

Bayesian methods typically do not rely on such strong regularity assumptions which make them appealing when dealing with notoriously complex models like mixtures. Inference on K can be carried out by using overfitted mixtures (i.e with $K > K_0$). Rousseau and Mengersen 2011 have indeed shown that for a specific choice of hyperparrameters on the mixture weights, the $K - K_0$ extra components are emptied, leaving only K_0 meaningful components. Alternatively, a natural strategy is to set a prior distribution on K resulting in the so-called Mixture of Finite Mixtures model (see for instance Miller and Harrison 2018 or Frühwirth-Schnatter et al. 2021 for a thorough review of this approach). However, Cai et al. 2021 show that the posterior estimate on the number of components K is not consistent when the mixture kernels f are misspecified, even when the prior distribution is allowed to depend on the data set size n.

The Bayesian paradigm offers a practical framework to address the issue of model choice thanks to the Bayes Factor (Jeffreys 1935), that conveniently allows to compare *any* two models $\mathfrak{M}_0$, $\mathfrak{M}_1$ through the identity

$$BF_{\mathfrak{M}_0,\mathfrak{M}_1} = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} \tag{2.1}$$

where $m_i(\boldsymbol{y})$, $i = 0, 1$, is the *marginal likelihood* of model $\mathfrak{M}_i$ (a.k.a *model evidence*) defined as the integral of the likelihood function against the prior distribution, that

is

$$m_i(\boldsymbol{y}) = \int_{\Theta_i} p_i(\boldsymbol{y}|\theta)\Pi_i(d\theta) \tag{2.2}$$

where p_i and Π_i denote the likelihood function and the prior distribution on Θ_i for model $\mathfrak{M}_i$, respectively. Straightforwardly, this quantity can also be viewed as the normalizing constant of the posterior distribution on θ. Note that there is typically no requirement that models be nested or have any particular structure for the Bayes Factor (2.1) to define a consistent model selection strategy. In fact, it can be shown for a large class of models that

$$BF_{\mathfrak{M}_0,\mathfrak{M}_1} \underset{n\to\infty}{\longrightarrow} +\infty$$

in $P_{\mathfrak{M}_0}-$ probability and interested readers are referred to Chib and Kuffner 2016 for a thorough review of consistency results for the Bayes Factor. Unfortunately, the marginal likelihood of the data is rarely easy to compute as integral (2.2) is usually intractable. This is the case for mixture models with more than 1 component. For instance, evaluating the marginal likelihood of a conjugate K-component mixture model for $\boldsymbol{y} = (y_1, \ldots, y_n)$ requires about $O(K^n)$ computations of analytical integrals, which is hardly achievable as soon as K and/or n become moderately large.

A common estimation strategy for $m_i(\boldsymbol{y})$ is through the Bayesian Information Criterion (BIC, Schwarz 1978) which makes use of a second-order Taylor's expansion of the unnormalized posterior density $\tilde{\pi}_i(\theta|\boldsymbol{y}) = p_i(\boldsymbol{y}|\theta)\pi_i(\theta)$ around its maximum θ^*. However, the derivation of BIC strongly relies on regularity and identifiability assumptions which do not hold for overfitted finite mixtures. To address this issue, Drton and Plummer 2017 suggests a modified *singular* BIC criterion. However, the effect of the prior distribution on $p_i(\boldsymbol{y})$ is neglected by BIC, which may not be desirable for Bayesian inference.

Therefore, it is common to resort to Monte Carlo methods to approximate the integral (2.2). Note that other approaches exist, in particular using variational methods. Interested readers can refer to Chérief-Abdellatif and Alquier 2018 for applications of variational Bayes techniques to mixture models.

Literature is rich with Monte Carlo methods for estimating the marginal likelihood of parametric models (see for instance Frühwirth-Schnatter 2004, Pajor 2017 or Llorente et al. 2020 for a thorough review of existing methods). However, the intrinsic complexity of mixture models calls for *ad hoc* algorithms or adaptations of existing methods. Among the difficulties posed by such models are the multi-modal nature of their posterior and the presence of a label-switching phenomenon (or lack thereof) when sampling from the latter, which will be explained and illustrated later in this chapter. Popular marginal likelihood estimators suitable for mixture models are the method of Chib 1995 or the adaptation of Bridge Sampling proposed by Frühwirth-Schnatter 2019, or, to some extent, Sequential Monte Carlo (SMC) methods (Del Moral et al. 2006).

However, if one wishes to retrieve an estimator with a reasonable variance, those methods get very computationally expensive as K, the number of mixture components, and/or n, the number of observations, increase. Values of K as small as 5 already represent a significant computational burden, stressing the need for

new estimators more suited to real-life applications where K needs not be small. For instance, some applications of topic models such as Latent Dirichlet Allocation (Blei et al. 2003) may involve a large number of latent topics. Large number of clusters may also naturally be found in big data applications such as in Ullah and Mengersen 2019 where as many as 15 groups are detected in the data. This chapter provides an assessment of Monte Carlo estimators of the model evidence (2.2) of Finite Mixture models, existing and new, and aims at identifying robust methods that scale well with K and n.

In this chapter, we first introduce the finite mixture model and derive all the necessary quantities to conduct Bayesian inference. We then describe and study the classical estimators of the marginal likelihood used for Finite Mixtures, that is the Arithmetic and Harmonic Mean Estimators, Chib's algorithm, Bridge Sampling and Sequential Monte Carlo and highlight their potential shortcomings. We then derive two novel algorithms, one of them making use of the latent partitioning structure induced by finite mixtures, which we believe can overcome the defects of the above-mentioned methods. Finally, we compare and assess in an empirical study the performances of all the algorithms under consideration.

2.2 The Finite Mixture model

In this section, we introduce the Finite Mixture model and provide basic Gibbs sampling strategies for simulating according to the posterior distribution. This work is mostly based on the book Frühwirth-Schnatter 2006, which we highly recommend to anyone wishing to read a thorough introduction to finite mixtures and Markov switching models.

2.2.1 Notation

- Data : $\boldsymbol{y} = (y_1, \dots, y_n)$, $n \geq 1$.
- Subset of data : $\boldsymbol{y}_{1:l} = (y_1, \dots, y_l)$ for $l \leq n$
- $\Pi(\cdot)$: generic notation of a prior distribution on some parameter ϑ.
- $\pi(\vartheta)$: Radon-Nikodym density of $\Pi(\cdot)$ with respect to some measure ν (the Lebesgue measure or the counting measure, depending on the context).
- $\Pi(\cdot|\boldsymbol{y})$: posterior distribution of some parameter ϑ given data $\boldsymbol{y}$.
- $\pi(\vartheta|\boldsymbol{y})$: Radon-Nikodym density of $\Pi(\cdot|\boldsymbol{y})$ with respect to some measure ν (the Lebesgue measure or the counting measure, depending on the context).

2.2.2 Model specification and posterior inference

A Finite Mixture model (FM) can be defined as the convex sum of K distributions which densities (with respect to some dominating measure) $f(\cdot|\theta_k)$, $k = 1, \dots, K$, are parametrized by some $\theta_k \in \Theta \subset \mathbb{R}^d$, for $d \in \mathbb{N} \setminus \{0\}$.
That is, given a collection of $n > 0$ independent and identically distributed ($i.i.d$) real random variables $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^{d \times n}$ assumed to arise from a finite mixture of K *emission* distributions $F(\cdot|\theta)$ with density with respect to the Lebesgue measure $f(\cdot|\theta)$, their joint distribution can be written as

$$p(y_1, \dots, y_n|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \varpi_k f(y_i|\theta_k). \tag{2.3}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}, \boldsymbol{\varpi})$ and $\boldsymbol{\varpi} = (\varpi_1, \dots, \varpi_K)$ belongs to the $K-1$-dimensional simplex Δ_{K-1}. In other words, for all $k = 1, \dots, K$, $\varpi_k \geq 0$ and $\varpi_1 + \dots + \varpi_K = 1$. From now on we shall denote by Ω_K the *complete* parameter space $\Theta^K \times \Delta_{K-1}$.
From a Bayesian perspective, it is necessary to elicit a prior distribution Π on Ω_K. It is common practice to assume that the *emission* parameter $\boldsymbol{\theta}$ and the vector of weights $\boldsymbol{\varpi}$ are independent *a priori*. The latter is typically assumed to follow a Dirichlet distribution, denoted by $\mathcal{D}(\cdot|\boldsymbol{\alpha})$, for some vector $\boldsymbol{\alpha} \in (0, \infty)^K$ and this shall be our prior choice thereafter. Hence, we define

$$\Pi_{\boldsymbol{\vartheta}}(d\boldsymbol{\vartheta}) := \left[\prod_{k=1}^{K} G_0(d\theta_k)\right] \mathcal{D}(d\boldsymbol{\varpi}|\alpha) \tag{2.4}$$

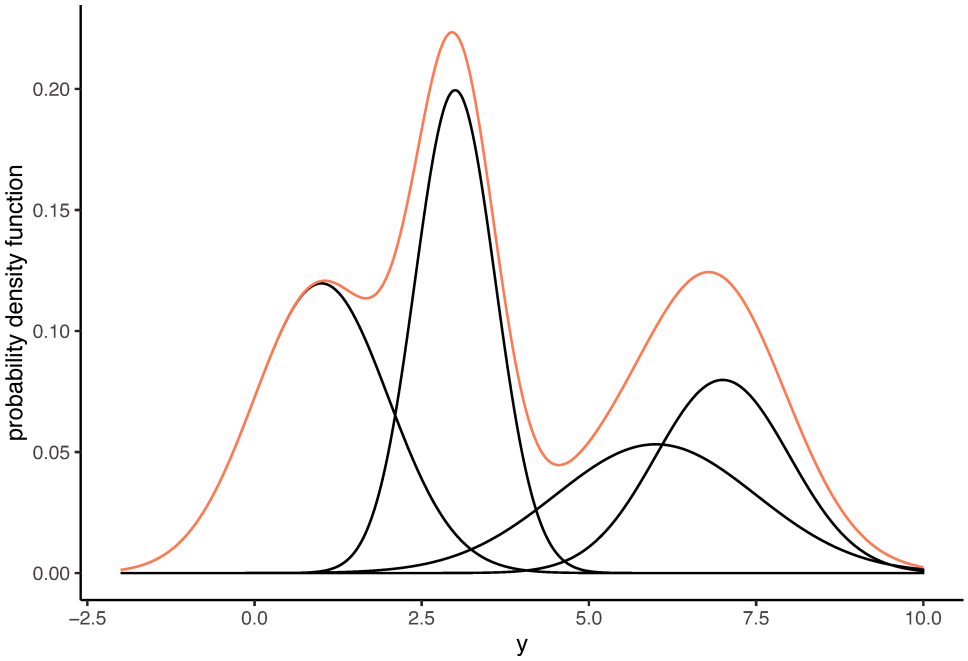


Figure 2.1: Mixture of four normal distributions given by $0.3\mathcal{N}(1,1)+0.3\mathcal{N}(3,0.6^2)+0.2\mathcal{N}(6,1.5^2)+0.2\mathcal{N}(7,1)$ (red line). The individual *weighted* densities are displayed in black.

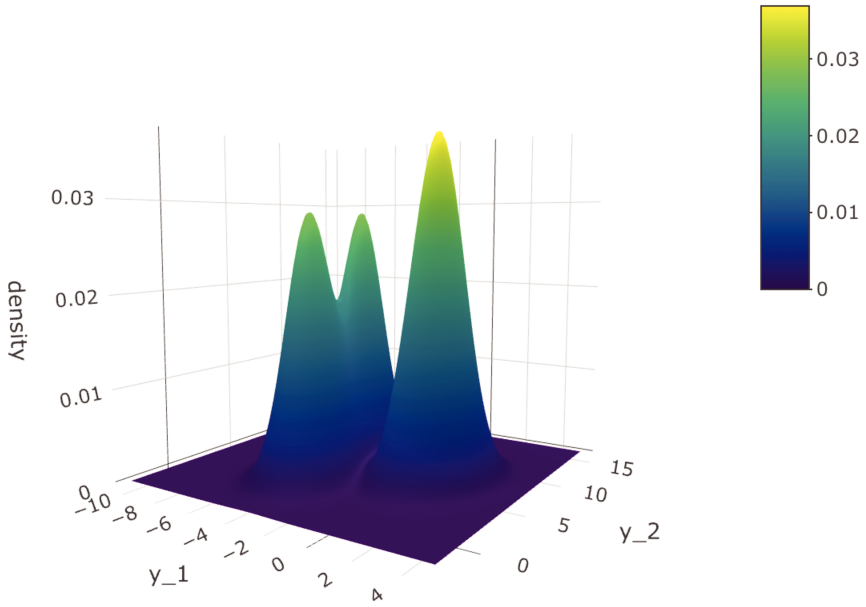


Figure 2.2: Bivariate mixture of 3 normal distributions $0.4\mathcal{N}\left(\begin{pmatrix}1\\6\end{pmatrix},\begin{pmatrix}1 & 0.2\\0.2 & 3\end{pmatrix}\right)+0.3\mathcal{N}\left(\begin{pmatrix}-4\\3\end{pmatrix},\begin{pmatrix}1 & -0.4\\-0.4 & 3\end{pmatrix}\right)+0.3\mathcal{N}\left(\begin{pmatrix}-4\\8\end{pmatrix},\begin{pmatrix}1 & -0.4\\-0.4 & 3\end{pmatrix}\right)$

where we have further assumed that the parameters $\theta_1, \dots, \theta_K$ are independent a priori and follow some distribution G_0 on Θ with density g_0 with respect to the Lebesgue measure.

Finite mixtures are key models to account for heterogeneity in a given population, as can be intuited from Figure 2.1, representing a simple mixture of 4 univariate normal distributions, or from Figure 2.2, for bivariate data. Therefore, a natural alternative specification of model (2.3) can be derived with the introduction of a latent *individual cluster membership* variable $\boldsymbol{z} = (z_1, \dots, z_n) \in \{1, \dots, K\}^n$, which indicates from which mixture component each observation y_i arises. Then sampling an observation y can be viewed as allocating a mixture component z with probability $\mathbb{P}(z = k) = \varpi_k$ for $k = 1, \dots, K$, before drawing y from the mixture component $f(\cdot|\theta_k)$. This generative view of mixture models can be summarized as

$$\begin{aligned} y_i|\boldsymbol{\theta}, z_i &\overset{i.i.d.}{\sim} f(.|\theta_{z_i}), && i = 1, \dots, n \\ \mathbb{P}(z_i = k|\boldsymbol{\varpi}) &= \varpi_k, && i = 1, \dots, n \\ \boldsymbol{\varpi} = (\varpi_1, \dots, \varpi_K) &\sim \mathcal{D}(\boldsymbol{\alpha}) \\ \theta_1, \dots, \theta_K &\overset{i.i.d.}{\sim} G_0 \end{aligned} \tag{2.5}$$

and allows for the definition of the complete data likelihood function (Frühwirth-Schnatter 2006)

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\vartheta}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\vartheta}, z_i) p(z_i|\boldsymbol{\vartheta}) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left(f(y_i|\theta_k)\varpi_k\right)^{\mathbb{1}(z_i=k)} \\ &= \prod_{k=1}^{K} \left(\prod_{i:z_i=k} f(y_i|\theta_k) \right) \left(\prod_{k=1}^{K} \varpi_k^{N_k(\boldsymbol{z})} \right) \end{aligned}$$

where $N_k(\boldsymbol{z}) = \sum_{i=1}^{N} \mathbb{1}(z_i = k)$, is the number of observations allocated to component $k = 1, \dots, K$.
Consequently, one can define the augmented posterior distribution using Bayes' formula by

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z}) \propto p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}). \tag{2.6}$$

The prior defined in (2.4) conveniently passes its independence property onto the posterior distribution (2.6) which factorizes as

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z}) = \prod_{k=1}^{K} \pi(\theta_k|\boldsymbol{y}, \boldsymbol{z}) \times \pi(\boldsymbol{\varpi}|\boldsymbol{z}). \tag{2.7}$$

where the posterior distribution on the weights $\boldsymbol{\varpi}$ can be easily identified with the Dirichlet distribution $\mathcal{D}(\cdot|\alpha_1 + N_1(\boldsymbol{z}), \dots, \alpha_K + N_K(\boldsymbol{z}))$ by noticing that

$$\pi(\boldsymbol{\varpi}|\boldsymbol{z}) \propto \prod_{k=1}^{K} \varpi_k^{N_k(z)} \times \varpi_k^{\alpha_k - 1}. \tag{2.8}$$

The posterior distribution on the parameters $\boldsymbol{\theta}$ can be derived up to a proportionality constant as

$$\pi(\theta_k|\boldsymbol{y},\boldsymbol{z}) \propto \prod_{i:z_i=k} f(y_i|\theta_k)g_0(\theta_k). \tag{2.9}$$

In the special case that g_0 is conjugate to f, the posterior on the parameters (2.9) has a closed form expression. The model is then called *conditionally conjugate* (Frühwirth-Schnatter 2006) since (2.9) belongs to the same parametric family as G_0, conditionally on the latent vector $\boldsymbol{z}$. We give below the example of a normally distributed mixture kernel F and its associated conjugate distribution G_0.

Example 1. If the mixture kernel F is chosen to be the normal distribution $\mathcal{N}(\cdot|\mu_k,\sigma_k^2)$ where $(\mu_k,\sigma_k^2)\in\mathbb{R}\times(0,\infty)$, then the associated conjugate prior distribution G_0 is the Normal-Inverse Gamma $\mathcal{NIG}(\cdot|\mu_0,\lambda,a,b)$ for $\mu_0\in\mathbb{R}$ and $\lambda,a,b>0$ defined as

$$\begin{aligned}\mathcal{NIG}(d\mu_k,d\sigma_k^2|\mu_0,\lambda,a,b) = &\frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma_k^2}}\frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma_k^2}\right)^{a+1}\\ &\times\exp\left\{-\frac{2b+\lambda(\mu_k-\mu_0)^2}{2\sigma^2}\right\}d\mu_k d\sigma_k^2.\end{aligned} \tag{2.10}$$

Indeed, let $\boldsymbol{y}_k := \{y_i : z_i = k\}$, the posterior distribution on the parameters $\theta_k := (\mu_k,\sigma_k^2)$ given $\boldsymbol{y}_k$ can be derived up to a proportionality constant as

$$\begin{aligned}\pi(\mu_k,\sigma_k^2|\boldsymbol{y},\boldsymbol{z}) &\propto \left(\frac{1}{\sqrt{\sigma_k^2}}\right)^{N_k(\boldsymbol{z})}\exp\left\{-\frac{\sum_{i:z_i=k}(y_i-\mu_k)^2}{2\sigma_k^2}\right\}\\ &\quad\times\frac{1}{\sqrt{\sigma_k^2}}\left(\frac{1}{\sigma_k^2}\right)^{a+1}\exp\left\{-\frac{2b+\lambda(\mu_k-\mu_0)^2}{2\sigma_k^2}\right\}\\ &\propto \frac{1}{\sqrt{\sigma_k^2}}\left(\frac{1}{\sigma_k^2}\right)^{N_k(\boldsymbol{z})/2+a+1}\exp\left\{-\frac{1}{2\sigma_k^2}\left[2\left(b+\frac{1}{2}\sum_{i:z_i=k}(y_i-\overline{\boldsymbol{y}_k})^2\right.\right.\right.\\ &\quad\left.\left.\left.+\frac{N_k(\boldsymbol{z})\lambda}{2(N_k(\boldsymbol{z})+\lambda)}(\overline{\boldsymbol{y}_k}-\mu_0)^2\right)+(\lambda+N_k(\boldsymbol{z}))\left(\mu_k-\frac{N_k(\boldsymbol{z})\overline{\boldsymbol{y}_k}+\lambda\mu_0}{N_k(\boldsymbol{z})+\lambda}\right)^2\right]\right\}\end{aligned}$$

Hence, the posterior distribution of (μ_k,σ_k^2) given $\boldsymbol{z}$ is again $\mathcal{NIG}(\mu_0',\lambda',a',b')$ where

$$\begin{cases}\mu_0' = (N_k(\boldsymbol{z})\overline{\boldsymbol{y}_k}+\lambda\mu_0)/(N_k(\boldsymbol{z})+\lambda)\\ \lambda' = \lambda+N_k(\boldsymbol{z})\\ a' = a+N_k(\boldsymbol{z})/2\\ b' = b+(1/2)\left[\sum_{i:z_i=k}(y_i-\overline{\boldsymbol{y}_k})^2+N_k(\boldsymbol{z})\lambda/(N_k(\boldsymbol{z})+\lambda)(\overline{\boldsymbol{y}_k}-\mu_0)^2\right]\end{cases}$$

Remark 2.1. *If*

$$\mu_k|\sigma_k^2,\mu_0,\lambda\sim\mathcal{N}(\mu_0,\sigma^2/\lambda)$$

and

$$\sigma_k^2|a,b \sim \mathcal{IG}(a,b)$$

where $\mathcal{IG}(a,b)$ *denotes the Inverse Gamma distribution with shape and scale* a *and* b*, then*

$$(\mu_k,\sigma_k^2) \sim \mathcal{NIG}(\mu_0,\lambda,a,b).$$

This alternative definition to the Normal-Inverse Gamma distribution comes in handy when deriving Gibbs sampling strategies to sample from the posterior distribution of $\theta_k = (\mu_k,\sigma_k^2)$*.*

However, knowledge about the allocations $\boldsymbol{z} = (z_1,\ldots,z_n)$ is usually not accessible and hence must be inferred along with the parameters $\boldsymbol{\theta}$ and the weights $\boldsymbol{\varpi}$. It turns out that the elements of $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{y}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta},\boldsymbol{\varpi})$ and that

$$\begin{aligned}\pi(z_i = k|y_i,\boldsymbol{\vartheta}) &= \frac{p(y_i|z_i=k,\boldsymbol{\vartheta})p(z_i=k|\boldsymbol{\vartheta})}{\sum_{j=1}^K p(y_i|z_i=j,\boldsymbol{\vartheta})p(z_i=j|\boldsymbol{\vartheta})}\\ &= \frac{f(y_i|\theta_k)\varpi_k}{\sum_{j=1}^K f(y_i|\theta_j)\varpi_j}\end{aligned} \tag{2.11}$$

The latent variable representation of mixture models (2.5) and the subsequent definition of the conditional posterior distributions on the parameters (2.7) and (2.8) and the allocation vector (2.11) are at the core of the Gibbs sampling strategy commonly used by practitioners for simulating $(\boldsymbol{\vartheta},\boldsymbol{z})$ from the posterior, which was introduced by Diebolt and Robert 1990. Algorithm 1 gives a pseudo-code implementation. Moreover, conjugacy allows to integrate parameters $\boldsymbol{\vartheta}$ out in order

Algorithm 1 : Gibbs sampler for conjugate finite mixture models

```
At step t,
for i = 1,...,n do
    Sample z_i^(t) | y_i, ϑ^(t-1) using (2.11);
end
for k = 1,...,K do
    Sample θ_k^(t) | y, z^(t) using (2.9)
end
Sample ϖ^(t) | z^(t) using (2.8)
```

to work solely with the allocation vector $\boldsymbol{z}$. Indeed, one can derive its likelihood as

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{z}) &= \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta})\Pi(d\boldsymbol{\theta}) \\
&= \int_{\Theta} \prod_{i=1}^{n} p(y_i|z_i,\boldsymbol{\theta})\Pi(d\boldsymbol{\theta}) \\
&= \prod_{k=1}^{K} \int_{\Theta} \prod_{i:z_i=k} p(y_i|\theta_k) G_0(d\theta_k) \\
&= \prod_{k=1}^{K} m_k(\boldsymbol{z}) \qquad (2.12)
\end{aligned}
$$

where $m_k(\boldsymbol{z}) := \int_{\Theta} \prod_{i:z_i=k} p(y_i|\theta_k) G_0(d\theta_k)$ for all $k = 1, \ldots, K$ is the marginal likelihood of the data allocated to component k. Note that a closed-form expression of this quantity is available thanks to the conjugacy of the prior and can be easily derived using Example 1 for the Normal Finite Mixture model, for instance.
Likewise, it is straightforward to compute the induced prior on $\boldsymbol{z}$ by integrating out the weights as follows

$$
\begin{aligned}
\pi(\boldsymbol{z}) &= \int_{\Delta_{K-1}} \pi(\boldsymbol{z}|\boldsymbol{\varpi})\mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}) \\
&= \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \int_{\Delta_{K-1}} \prod_{k=1}^{K} \varpi_k^{N_k(\boldsymbol{z})+\alpha_k-1} d\varpi_1 \ldots d\varpi_K \\
&= \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\prod_{k=1}^{K}\Gamma(N_k(\boldsymbol{z})+\alpha_k)}{\Gamma\left(n+\sum_{k=1}^{K}\alpha_k\right)\prod_{k=1}^{K}\Gamma(\alpha_k)}. \qquad (2.13)
\end{aligned}
$$

Therefore, it is natural to work with the posterior on the allocations given by

$$\pi(\boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}). \qquad (2.14)$$

Sampling from (2.14) can be done with Gibbs sampling by writing the posterior full conditionals for all $i = 1, \ldots, n$,

$$\pi(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{z}_{-i}, \{z_i = k\})\pi(\boldsymbol{z}_{-i}, \{z_i = k\})}{\sum_{j=1}^{K} p(\boldsymbol{y}|\boldsymbol{z}_{-i}, \{z_i = j\})\pi(\boldsymbol{z}_{-i}, \{z_i = j\})}$$

where $\boldsymbol{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$. The collapsed Gibbs sampler (Algorithm 2), as first described by Chen and Liu 1996, makes use of this representation to sample the allocation vector from the posterior distribution, without simulating the parameters $\boldsymbol{\vartheta}$.

Posterior simulations, as provided by Algorithms 1 and 2, are at the core of most Monte Carlo methods aiming at estimating the marginal likelihood of a Finite Mixture model, as we shall see in this chapter. In the next section we briefly introduce the concept of non-identifiability and show how it makes the posterior distribution invariant under permutations of the model parameters.

Algorithm 2 : Collapsed Gibbs sampler on the allocations

1 At step t,
2 **for** $i = 1, \ldots, n$ **do**
3 | Sample a new allocation $z_i^{(t)} | \boldsymbol{z}_{-i}^{(t-1)}, \boldsymbol{y}$ using 8
4 **end**

2.2.3 Non-identifiability of the model : posterior permutation invariance and label-switching

Identifiability is a desirable feature of parametric models in order to conduct meaningful inference of the underlying parameters. More precisely, a statistical model $\mathfrak{M}_\lambda = \{\lambda \in \Lambda : P_\lambda\}$ on some sample space $\mathcal{Y}$ is identifiable *iff*

$$p(\boldsymbol{y}|\lambda_1) = p(\boldsymbol{y}|\lambda_2) \text{ for almost all } \boldsymbol{y} \in \mathcal{Y} \Rightarrow \lambda_1 = \lambda_2$$

for some $\lambda_1, \lambda_2 \in \Lambda$.
It is clear that finite mixture models do not satisfy this condition. Indeed, consider a $K-$component finite mixture with some mixture kernel $f(\cdot|\theta)$ and let $\boldsymbol{\vartheta}^*$ be a permuted version of some $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi})$ in the complete parameter space Ω_K. Then

$$p(y|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \varpi_k f(y|\theta_k) = \sum_{k=1}^{K} \varpi_k^* f(y|\theta_k^*) = p(y|\boldsymbol{\vartheta}^*)$$

while $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}^*$.
This permutation-invariance property of the likelihood function propagates to the posterior distribution, provided the prior is also permutation-invariant. Our choice of prior so far satisfies this invariance property provided that $\boldsymbol{\alpha}$, the hyperparameter of the Dirichlet prior on the weights $\boldsymbol{\varpi}$ is such that $\alpha_1 = \cdots = \alpha_K$.

In such cases, the permutation-invariance of the posterior distribution can easily be deduced from the fact that

$$p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) = p(\boldsymbol{y}|\boldsymbol{\vartheta}^*)\pi(\boldsymbol{\vartheta}^*)$$

which implies that

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{y}) = \pi(\boldsymbol{\vartheta}^*|\boldsymbol{y}) \tag{2.15}$$

for all permutation $\boldsymbol{\vartheta}^*$ of $\boldsymbol{\vartheta} \in \Omega_K$. This implies that the posterior distributions on the parameters $\boldsymbol{\vartheta}$ has $K!$ equiprobable modes, as illustrated in Figure 2.3. In this Figure, we observe a so called balanced *label-switching* phenomenon. That is, the posterior samples are such that θ_1, the parameter associated to the 'first' mixture component, is sampled equally from both the neighborhoods of the true parameter $\boldsymbol{\theta}_0$ and its permuted version $\boldsymbol{\theta}_0^*$. This is a desirable property of MCMC samplers. However, as K gets large or when the data has relatively distant modes, it typically becomes increasingly difficult for samplers to visit equally all $K!$ posterior modes. Posterior invariance is a challenge for practitioners since equation (2.15) hinders estimation of the parameters a posteriori. For instance, it implies that the classical

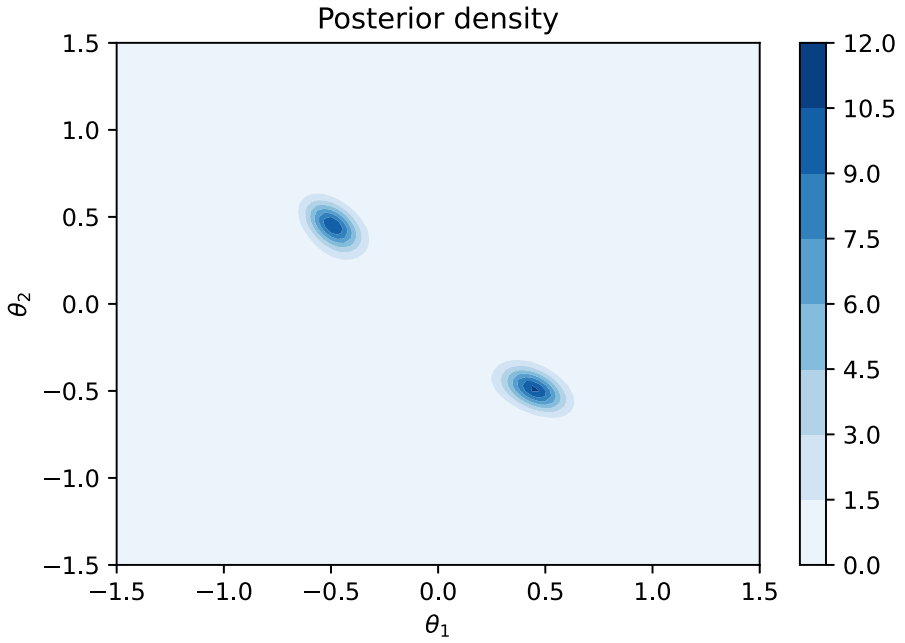


Figure 2.3: Density plot of 200 000 posterior simulations from a 2-component mixture of Normal distributions, with true paramerers $\boldsymbol{\theta}_0 = (0.5, -0.5)$ and equal weights $\boldsymbol{\varpi}_0 = (0.5, 0.5)$

posterior mean estimator of the component specific θ_k is such that $\mathbb{E}[\theta_k|\boldsymbol{y}] = \mathbb{E}[\theta_{k'}|\boldsymbol{y}]$, for all $k, k' \in \{1, \ldots, K\}$. Note that choosing an asymmetric prior, while effectively making the various modes not equally likely a posteriori, is very unlikely to reduce the posterior to a unimodal distribution. Alternatively, practitioners sometimes restrict the parameter space a priori by forcing $\theta_1 < \cdots < \theta_K$, for instance, but this often leads to impractical posterior sampling (Celeux et al. 2000).

As we shall see in the next section, the lack of identifiability of finite mixtures and the multimodality of their posterior distribution severely hinders the estimation of the model evidence.

2.3 Classical estimators and their shortcomings

In this section we review the most popular algorithms for estimating the marginal likelihood of finite mixtures and highlight how they struggle tackling the challenging structure of mixture models. In particular, we explain why some of them fail at giving reliable estimates at a reasonable computational cost as soon as the number of mixture components K is bigger than 5.

2.3.1 Arithmetic and harmonic mean estimators

We begin our review with two naive estimators, namely the Arithmetic Mean estimator (AME) and the Harmonic Mean Estimator (HME), and explain how, despite their convenient simplicity, they generally produce poor estimates of $m(\boldsymbol{y})$ for finite mixtures.

The AME is derived quite straightforwardly by noticing that

$$m(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\vartheta})\Pi(d\boldsymbol{\vartheta}) = \mathbb{E}_{\Pi}\left[p(\boldsymbol{y}|\boldsymbol{\vartheta})\right]$$

which implies that the following estimator

$$\widehat{m}_{AME}(\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t)}) \tag{2.16}$$

where $\{\boldsymbol{\vartheta}^{(t)}\}_{t=1}^{T} \overset{i.i.d}{\sim} \Pi$, is unbiased for $m(\boldsymbol{y})$. It is assumed that one can generate $i.i.d$ samples from the prior distribution, which is often the case in practice.
By its simplicity, this estimator is often used as a first approach to estimating the marginal likelihood of a model. However, whenever the support of the prior is much more spread out than that of the posterior, as is often the case when the sample size n grows, most prior samples will result in a near-zero likelihood evaluation, while only a very small portion of them contributes to the mean in (2.16). This means that in some situations, a prohibitively large number of prior simulations is necessary to get a reliable estimate of the marginal likelihood through the AME.

A corrected version of the AME was proposed by Pajor 2017, by constraining the prior simulations to a high-posterior region A and compensating by dividing the estimator (2.16) by an estimate of $\Pi(A|\boldsymbol{y})$. This method is not really suited to finite mixtures in that the posterior distribution is heavily multimodal, as discussed in Section 2.2.3, making the choice of A difficult.

The Harmonic Mean Estimator, first introduced by Newton and Raftery 1994 is

built upon the identity

$$\begin{aligned} 1 &= \int_{\Theta} \pi(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\ \Leftrightarrow 1 &= m(\boldsymbol{y}) \int_{\Theta} \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{p(\boldsymbol{y}|\boldsymbol{\vartheta})m(\boldsymbol{y})} d\boldsymbol{\vartheta} \\ \Leftrightarrow m(\boldsymbol{y}) &= \left[\int_{\Theta} \frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{\vartheta})} \right]^{-1} \\ \Leftrightarrow m(\boldsymbol{y}) &= \left\{ \mathbb{E}_{\Pi(\boldsymbol{\vartheta}|\boldsymbol{y})} \left[\frac{1}{p(\boldsymbol{y}|\boldsymbol{\vartheta})} \right] \right\}^{-1} \end{aligned}$$

which leads to the estimator

$$\widehat{m}_{HME}(\boldsymbol{y}) = \left[\frac{1}{T} \sum_{t=1}^{T} \frac{1}{p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t)})} \right]^{-1}$$

where $\{\boldsymbol{\vartheta}^{(t)}\}_{t=1}^{T} \overset{i.i.d}{\sim} \pi(\boldsymbol{\vartheta}|\boldsymbol{y})$ This estimator has the advantage of making use of posterior simulations instead of samples from the prior. However, this estimator suffers from many shortcoming. First, the inverse of the likelihood function might be really large for some points in the parameter space, leading to a potentially infinite variance. Second, a simple application of Jensen's inequality shows that the HME overestimates the marginal likelihood. Finally, as pointed out by Lenk 2009, it suffers from a so called *simulation pseudo-bias*. This phenomenon is due to the fact that most posterior simulations tend to concentrate on a subset $\tilde{\Theta}$ of the parameter space Θ such that $\Pi(\tilde{\Theta}|\boldsymbol{y}) = 1 - \varepsilon$ for some small $\varepsilon > 0$. This is in part due to the absolute zero of the computer preventing the Markov Chain Monte Carlo (MCMC) algorithm targeting the posterior to visit certain areas of the parameter space with very low posterior probability. While this is not really a problem for parameter inference, it is troublesome for marginal likelihood estimation. Indeed, one can write

$$\begin{aligned} m(\boldsymbol{y}) &= \int_{\tilde{\Theta}} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} + \int_{\tilde{\Theta}^c} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\ &= \Pi(\tilde{\Theta}) \int_{\tilde{\Theta}} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\tilde{\Theta}) d\boldsymbol{\vartheta} + m(\boldsymbol{y})\Pi(\tilde{\Theta}^c|\boldsymbol{y}) \\ &= \Pi(\tilde{\Theta}) \int_{\tilde{\Theta}} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\tilde{\Theta}) d\boldsymbol{\vartheta} + m(\boldsymbol{y})\varepsilon \\ &= \frac{\Pi(\tilde{\Theta})}{1-\varepsilon} \int_{\tilde{\Theta}} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\tilde{\Theta}) d\boldsymbol{\vartheta} \end{aligned}$$

Since MCMC posterior simulations are concentrated in the subset $\tilde{\Theta}$, then the HME actually estimates the quantity $\int_{\tilde{\Theta}} p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\tilde{\Theta}) d\boldsymbol{\vartheta}$ but is missing a factor $\Pi(\tilde{\Theta})/(1-\varepsilon)$. In practice $\Pi(\tilde{\Theta})$ might be much smaller than one, in particular when the prior is very diffuse, as soon as the likelihood concentrates on a small subset of Θ. This leads to an overestimation of the log-marginal likelihood by $\log \Pi(\tilde{\Theta})$.

In favorable scenarios, the AME and HME can yield accurate estimates of the

marginal likelihood, but it is simple to see that they struggle when confronted to the complex structure of mixture models, which is highlighted in the simulations of Section 2.5. Therefore, tailor-made methods have been developed over the years in order to tackle this issue and we next review Chib's estimator, that is usually deemed to be the *gold standard* for estimating the marginal likelihood of finite mixture models.

2.3.2 Chib's estimator

Chib's estimator was first introduced in Chib 1995 as a convenient way to estimate the marginal likelihood of a model directly from the output of a MCMC algorithm targeting the posterior distribution. It builds upon the simple identity

$$m(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta}_0)\pi(\boldsymbol{\vartheta}_0)}{\pi(\boldsymbol{\vartheta}_0|\boldsymbol{y})} \tag{2.17}$$

which holds for all value of $\boldsymbol{\vartheta}_0 \in \Omega_K$ by Bayes' formula.

Therefore it is enough to be able to evaluate (2.17) at a suitably chosen point $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\varpi}_0)$ in the parameter space. The likelihood function and the prior density are available in closed form, hence it is easy to compute $p(\boldsymbol{y}|\boldsymbol{\vartheta}_0)$ and $\pi(\boldsymbol{\vartheta}_0)$. However, the posterior density is intractable and requires careful estimation. The idea of Chib 1995 is to use the fact that the conditional posterior density $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z})$ has a simple formulation given in (2.7), given a conditionally conjugate prior G_0. Since

$$\pi(\boldsymbol{\vartheta}_0|\boldsymbol{y}) = \int \pi(\boldsymbol{\vartheta}_0|\boldsymbol{y}, \boldsymbol{z})\pi(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z},$$

the posterior density can be approached at point $\boldsymbol{\vartheta}_0$ through the *Rao-Blackwellized* unbiased estimator

$$\widehat{\pi}(\boldsymbol{\vartheta}_0|\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} \pi(\boldsymbol{\vartheta}_0|\boldsymbol{y}, \boldsymbol{z}^{(t)}) \tag{2.18}$$

where $\{\boldsymbol{z}^{(t)}\}$ is a sample of the posterior distribution $\pi(\boldsymbol{z}|\boldsymbol{y})$.
The estimate (2.18) is then plugged into (2.17) yielding Chib's estimator of the marginal likelihood

$$\widehat{m}_{Chib}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta}_0)\pi(\boldsymbol{\vartheta}_0)}{\widehat{\pi}(\boldsymbol{\vartheta}_0|\boldsymbol{y})}. \tag{2.19}$$

Note that for the posterior estimate (2.18) to have good variance properties, it is usually recommended to choose some high posterior density point. Typically, choosing the maximum a posteriori (MAP) estimator from the Gibbs output is a good strategy since the likelihood and the prior can be easily evaluated. Algorithm 3 gives a pseudo-code implementation of Chibs estimator.

Unfortunately, as pointed out by Neal 1999, despite being theoretically valid, Chib's method is severely affected by the multimodality of the mixture posterior distribution. Indeed, it is unusual for the Gibbs sampler to visit equally all $K!$ modes of the posterior distribution. In fact, it is not rare for it to get *stuck* in only one modal configuration as illustrated in Figure 2.4, where the example of a two-component mixture is given. In this case, (2.18) overestimates the posterior

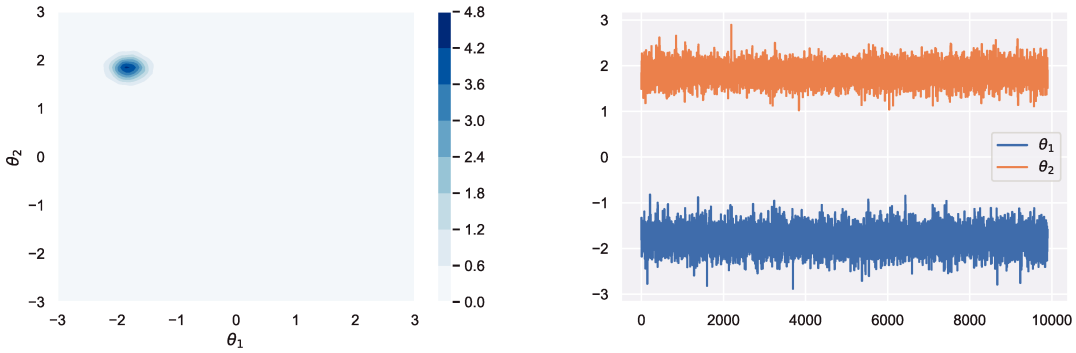
Algorithm 3 : Chib's algorithm for conditionally conjugate finite mixtures

Input : $\{\boldsymbol{\vartheta}^{(t)}, \boldsymbol{z}^{(t)}\}$, $t = 1; \ldots, T$ from the posterior $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}|\boldsymbol{y})$

1 $\boldsymbol{\vartheta}_0 = \arg\max_t p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t)})\pi(\boldsymbol{\vartheta}^{(t)})$; `// Identify the MAP estimate`

2 Compute $\widehat{\pi}(\boldsymbol{\vartheta}_0|\boldsymbol{y})$ using (2.18)

3 Return $\widehat{m}_{Chib}(\boldsymbol{y})$ using (2.19)



(a) Contour plot (b) Trace plots

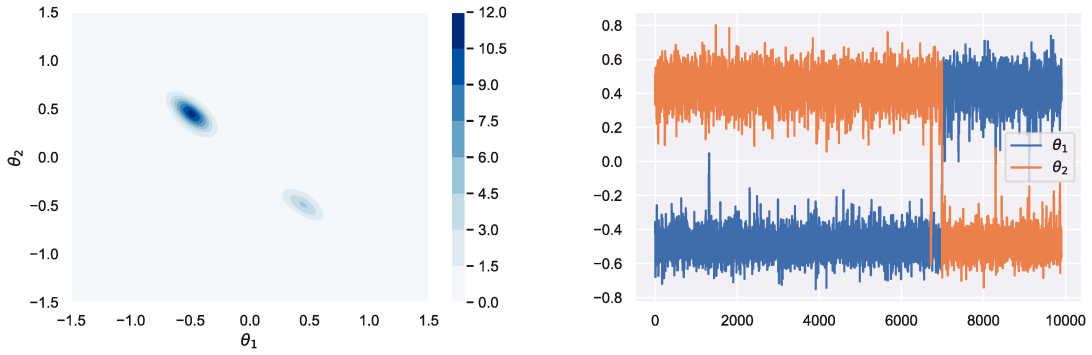
Figure 2.4: Contour and trace plots of 10 000 simulations from the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of a 2-component mixture of Normal distributions fitted to data arising from the mixture $0.4 \times \mathcal{N}(-2, 0.1^2) + 0.6 \times \mathcal{N}(2, 0.1^2)$

density at $\boldsymbol{\vartheta}_0$. To understand this phenomenon, note that, typically, for a fixed t,

$$\pi(\boldsymbol{\vartheta}^{(t)}|\boldsymbol{y}, \boldsymbol{z}^{(t)}) \geq \pi(\boldsymbol{\vartheta}^{(t)}|\boldsymbol{y}, \boldsymbol{z}^{(\boldsymbol{t})^*}) \tag{2.20}$$

where $\boldsymbol{z}^{(t)^*} = (\sigma(z_1), \ldots, \sigma(z_n))$, σ being a permutation of the set $\{1, \ldots, K\}$. This is because $\boldsymbol{z}^{(t)}$ is the allocation configuration that was used in the Gibbs sampler (Algorithm 1) to generate $\boldsymbol{\vartheta}^{(t)}$. In fact, the right hand side of equation (2.20) can be very close to zero whenever the parameters are well estimated. Hence, in the extreme case of Figure 2.4 where only one mode is visited, the posterior density will be estimated exclusively with values of the allocation vector $\boldsymbol{z}$ which labeling matches the one of $\boldsymbol{\vartheta}_0$. However, since all the permutations of $\boldsymbol{z}^{(t)}$ are equally likely a posteriori, (2.18) over-estimates the true posterior density by approximately a factor $K!$. This leads in turn to an under-estimation of the marginal likelihood by a factor $1/K!$.

An easy fix would be to directly correct the bias by multiplying (2.19) by $K!$. However, this only applies to scenarios like Figure 2.4 but in practice, the Gibbs sampler may visit other modes, but not necessarily all of them, nor evenly. Hence the bias factor might be less than $K!$ but not equal to 1 either. Figure 2.5 illustrates such a scenario where the so called *label switching* phenomenon takes place but the two posterior modes are not equally visited by the Gibbs sampler. A more elaborate solution was proposed by Berkhof et al. 2003 and Marin and Robert 2008 and consists in averaging over all possible permutations of $\boldsymbol{z}^{(t)}$ for $t = 1, \ldots, T$, yielding the following estimator



(a) Contour plot

(b) Trace plots

Figure 2.5: Contour and trace plots of 10 000 simulations from the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of a 2-component mixture of Normal distributions fitted to data arising from the mixture $0.5 \times \mathcal{N}(-0.5, 0.2^2) + 0.5 \times \mathcal{N}(0.5, 0.2^2)$

$$\widehat{m}_{ChibPerm}(\boldsymbol{y}) = \frac{1}{TK!} \sum_{t=1}^{T} \sum_{\sigma \in \mathfrak{S}(\{1,\ldots,K\})} \pi(\boldsymbol{\vartheta}_0|\boldsymbol{y}, \sigma(z_1^{(t)}), \ldots, \sigma(z_n^{(t)})) \tag{2.21}$$

where $\mathfrak{S}(\{1,\ldots,K\})$ denotes the set of permutations of the $\{1,\ldots,K\}$. Plugging this quantity in (2.17) yields the *permutation Chib's estimator*. A more computationally efficient version of (2.21) is obtained by permuting the parameters instead of the allocations.

$$\widehat{m}_{ChibPerm}(\boldsymbol{y}) = \frac{1}{TK!} \sum_{t=1}^{T} \sum_{\sigma \in \mathfrak{S}(\{1,\ldots,K\}} \pi((\vartheta_{0\sigma(1)}, \ldots, \vartheta_{0\sigma(K)})|\boldsymbol{y}, \boldsymbol{z}^{(t)}) \tag{2.22}$$

Algorithm 4 : Permutation Chib's algorithm for conditionally conjugate finite mixtures

Input : $\{\boldsymbol{\vartheta}^{(t)}, \boldsymbol{z}^{(t)}\}$, $t = 1;\ldots,T$ from the posterior $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}|\boldsymbol{y})$

1 $\boldsymbol{\vartheta}_0 = \arg\max_t p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t)})\pi(\boldsymbol{\vartheta}^{(t)})$; `// Identify the MAP estimate`

2 Return $\widehat{m}_{ChibPerm}(\boldsymbol{y})$ using (2.22)

This solution is convenient but can get very computationally expensive as soon as the number of components K gets moderately large - values as small as 5 already represent a significant computational burden.

A trick suggested by Marin and Robert 2008 and Lee and Robert 2016 is to sample at random 100 permutations in $\mathfrak{S}(\{1,\ldots,K\})$ for such values of K in order to keep computational time below a reasonable threshold. Unfortunately, 100 random permutations should not be enough to keep a reasonable variance for Chib's estimator when $K! >> 100$.

2.3.3 Bridge Sampling

Bridge sampling is a very popular generalization of Importance sampling that was introduced by Meng and Wong 1996 as a way to directly estimate the ratio of normalizing constants of two distributions. An immediate application is the estimation of the Bayes Factor between two models, for instance. If one of those distributions has a known normalizing constant, Bridge sampling can be used to derive an estimate of the marginal likelihood of the other distribution.
Introducing $q(\boldsymbol{\vartheta})$, an importance distribution for $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$ supported on Ω_K and $\alpha(\boldsymbol{\vartheta})$, a positive function such that

$$0 < \int_{\Omega_K} \alpha(\boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}|\boldsymbol{y}) < \infty$$

then one can notice that

$$\begin{aligned} \frac{\mathbb{E}_{q(\boldsymbol{\vartheta})}\left[\alpha(\boldsymbol{\vartheta}) p(\boldsymbol{y}|\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta})\right]}{\mathbb{E}_{\pi(\boldsymbol{\vartheta}|\boldsymbol{y})}\left[\alpha(\boldsymbol{\vartheta}) q(\boldsymbol{\vartheta})\right]} &= \frac{\int_{\Omega_K} \alpha(\boldsymbol{\vartheta}) p(\boldsymbol{y}|\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}{\int_{\Omega_K} \alpha(\boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{m(\boldsymbol{y})} d\boldsymbol{\vartheta}} \\ &= m(\boldsymbol{y}) \end{aligned} \tag{2.23}$$

Identity (2.23) naturally leads to the following Bridge Sampling estimator of the marginal likelihood

$$\widehat{m}_{BS}(\boldsymbol{y}) = \frac{1}{T_1} \sum_{t=1}^{T_1} \alpha(\boldsymbol{\vartheta}^{(t,1)}) p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t,1)}) \pi(\boldsymbol{\vartheta}^{(t,1)}) \Big/ \frac{1}{T_2} \sum_{t=1}^{T_2} \alpha(\boldsymbol{\vartheta}^{(t,2)}) q(\boldsymbol{\vartheta}^{(t,2)}) \tag{2.24}$$

where $\{\boldsymbol{\vartheta}^{(t,1)}\}_{t=1}^{T_1}$ are simulations from the importance distribution $q(\boldsymbol{\vartheta})$ whereas $\{\boldsymbol{\vartheta}^{(t,2)}\}_{t=1}^{T_2}$ are simulations from the posterior $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$.
The choice of $\alpha(\boldsymbol{\vartheta})$ is crucial to guarantee a reasonable variance for estimator (2.24). The optimal $\alpha(\boldsymbol{\vartheta})$ that minimizes the variance is derived by Meng and Wong 1996 as

$$\alpha^*(\boldsymbol{\vartheta}) = \left[T_1 q(\boldsymbol{\vartheta}) + T_2^* \pi(\boldsymbol{\vartheta}|\boldsymbol{y})\right]^{-1}$$

where T_2^* denotes the Effective Sample Size (ESS) of the posterior simulations, as they most likely arise from a MCMC algorithm and hence are not $i.i.d.$ The optimal α^* given above is useless in practice as the posterior distribution cannot be evaluated point-wise. A trick is to define recursively the Bridge Sampling estimator as

$$\widehat{m}_{BS,l}(\boldsymbol{y}) = \frac{\frac{1}{T_1} \sum_{t=1}^{T_1} \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t,1)})\pi(\boldsymbol{\vartheta}^{(t,1)})}{T_1 q(\boldsymbol{\vartheta}^{(t,1)}) + T_2^* \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t,1)})\pi(\boldsymbol{\vartheta}^{(t,1)})}{m_{BS,l-1}(\boldsymbol{y})}}}{\frac{1}{T_2} \sum_{t=1}^{T_2} \frac{q(\boldsymbol{\vartheta}^{(t,2)})}{T_1 q(\boldsymbol{\vartheta}^{(t,2)}) + T_2^* \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t,2)})\pi(\boldsymbol{\vartheta}^{(t,2)})}{m_{BS,l-1}(\boldsymbol{y})}}} \tag{2.25}$$

where $p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})/\widehat{m}_{BS,l-1}(\boldsymbol{y})$ is used as an approximation to $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$ at iteration l. At the first iteration, $m_{BS,0}(\boldsymbol{y})$ can be chosen to be a simple arithmetic mean estimation of the marginal likelihood.

As l grows, the optimal Bridge sampling estimator is retrieved and defined as

$$\widehat{m}_{BS}(\boldsymbol{y}) = \lim_{l\to\infty} \widehat{m}_{BS,l}(\boldsymbol{y}).$$

In practice, one should repeat recursion (2.25) until a stopping criterion is met, such as $|m_{BS,l}(\boldsymbol{y}) - m_{BS,l-1}(\boldsymbol{y})| \leq \epsilon$, for some small $\epsilon > 0$.

Another crucial choice for a successful Bridge sampling estimation of the marginal likelihood is that of the distribution $q(\boldsymbol{\vartheta})$. Frühwirth-Schnatter 2019 provides a comprehensive review on how to apply the Bridge Sampling framework to finite mixtures. As can be expected, the difficulty lies in deriving a good importance density $q(\boldsymbol{\vartheta})$ that yields a good approximation to the notoriously complex posterior distribution of a finite mixture model. In particular, q must have support on the $K!$ posterior modes. In the same vein as the permutation Chib's estimator (2.21) described in the previous section, Frühwirth-Schnatter 2019 proposes the following choice for $q(\boldsymbol{\vartheta})$,

$$q(\boldsymbol{\vartheta}) = \frac{1}{K!T_0} \sum_{t=1}^{T_0} \sum_{\sigma\in\mathfrak{S}(\{1,\ldots,k\})} \pi(\boldsymbol{\vartheta}|\sigma(\boldsymbol{z}^{(t,0)}), \boldsymbol{y})$$

where $T_0 \ll T_1$ and $\{\boldsymbol{z}^{(t,0)}\}_{t=1}^{T_0}$ is drawn with replacement from the output $\{\boldsymbol{z}^{(t,1)}\}_{t=1}^{T_1}$ of the Markov chain targeting the posterior.
Notice that q is simply a mixture of $K!T_0$ distributions with equal weights. Hence, sampling from q can simply be done by first drawing uniformly one of the T_0 allocation vector $\boldsymbol{z}^{(t,0)}$, then drawing a permutation σ and finally sampling $\boldsymbol{\vartheta}$ given $(\sigma(z_1), \ldots, \sigma(z_n))$ and $\boldsymbol{y}$ using (2.7).
Although T_0 is a smaller number than T_1, it is clear that this estimator suffers from the same type of computational shortcomings as the permutation Chib's estimator. Computing estimator (2.25) comes at the cost of evaluating $T_0 \times K! \times (T_1 + T_2)$ times the augmented posterior $\pi(\boldsymbol{\vartheta}|\boldsymbol{z}, \boldsymbol{y})$ distribution.

Algorithm 5 : Bridge Sampling Algorithm for conditionally conjugate finite mixtures

Input : $\{\boldsymbol{\vartheta}^{(t,1)}\}$, $t = 1; \ldots, T_1$ i.i.d from the importance distribution $q(\boldsymbol{\vartheta})$
$\{\boldsymbol{\vartheta}^{(t,2)}\}$, $t = 1; \ldots, T_2$ from the posterior $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}|\boldsymbol{y})$
Tolerance $\epsilon > 0$

`/* Initialize the sequence` $m_{BS,l}$ `*/`

Set $\widehat{m}_{BS,0}(\boldsymbol{y}) = \widehat{m}_{AME}(\boldsymbol{y})$ using (2.16)
Set $l = 0$
do
  $l = l + 1$
  Compute $\widehat{m}_{BS,l}$ using (2.25)
while $|\widehat{m}_{BS,l} - \widehat{m}_{BS,l-1}| > \epsilon$;
Return $m_{BS,l}$

2.3.4 Sequential Monte Carlo

Sequential Monte Carlo (SMC) methods are a broad class of algorithms that successively perform importance sampling and resampling steps on a sequence of instrumental distributions in order to obtain a sample from a distribution of interest. They are particularly popular for state space models such as Hidden Markov Models (HMM) for instance (see e.g Kantas et al. 2015), although their range of application is much broader. In the context of Bayesian inference, the goal is to design a sequence of distributions $\{\pi_l\}_{l=0}^L$, where π_0 typically is the prior distribution and π_L the posterior distribution of interest, $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. Starting with an initial sample $\{\boldsymbol{\vartheta}_t^{(0)}\}_{t=1}^T$ from π_0, the algorithm successively reweights the particles $\boldsymbol{\vartheta}_t$ through the importance step

$$w_t^{(l)} \propto \frac{\pi_l(\boldsymbol{\vartheta}_t^{(l-1)})}{\pi_{l-1}(\boldsymbol{\vartheta}_t^{(l-1)})} \tag{2.26}$$

where initially $w_t^{(0)} = \pi_0(\boldsymbol{\vartheta}_t^{(0)})$, yielding the weighted sample $\{\boldsymbol{\vartheta}_t^{(l)}, w_t^{(l)}\}$. It is then common practice to resample with replacements the particles $\{\boldsymbol{\vartheta}_t\}_t$ according to their normalized weights $W_t^{(l)} = w_t^{(l)} / \sum_{j=1}^T w_j^{(l)}$ yielding a new sample with equal normalized weights $\{\tilde{\boldsymbol{\vartheta}}_t^{(l)}, 1/T\}$. This resampling step, while enabling the deletion of unlikely particles, might lead to a so called *particle degeneracy* as the number of unique values among the sample may become low. To enforce diversity among the set of particles, a *mutation* step is performed in which each $\boldsymbol{\vartheta}_t^{(l)}$ is moved through a π_l-invariant MCMC kernel $K_l(\cdot, d\boldsymbol{\vartheta})$. In practice, the mutation kernel is applied several times on each particle.

Equation (2.26) above stresses the importance of choosing successive distributions π_t and π_{t+1} as not being too dissimilar in order to ensure good variance properties. Otherwise, the last set of particles $\{\boldsymbol{\vartheta}_t^{(L)}, w_t^{(L)}\}$ might be a very poor approximation to the distribution of interest π_L. Chopin 2002 adopts a data tempering approach, in which the sequence $\{\pi_l\}_l$ is chosen to be $\{\pi(\boldsymbol{\vartheta}|\boldsymbol{y}_{1:a_l})\}_{l=1}^L$ where $\boldsymbol{y}_{1:a_l} = (y_1, \dots, y_{a_l})$ and $a_1 < a_2 < \cdots < a_L = n$. This approach is computationally interesting in that it does not require the evaluation of the likelihood function on the whole dataset $\boldsymbol{y}$ at each iteration. A possible limitation is the underlying assumption that $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}_{1:a_l})$ and $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}_{1:a_{l+1}})$ are likely to be similar, which might heavily depend on the construction of the data batches $\{\boldsymbol{y}_{1:a_l}\}_l$, especially for finite mixtures.

We describe here another approach based on a sequence of so called *tempered* posteriors

$$\pi_l(\boldsymbol{\vartheta}) = \pi_l(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\vartheta}) p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_l}$$

in which the likelihood is raised to a temperature λ_l such that $\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_{L-1} < \lambda_L = 1$. For suitably chosen temperatures, one can effectively bridge the gap from the prior to the posterior distribution.

On top of being an alternative to classical MCMC schemes, the SMC framework provides an immediate estimate of the marginal likelihood. Indeed, let

$$Z_l := \int \pi(\boldsymbol{\vartheta}) p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_l} d\boldsymbol{\vartheta}$$

be the normalizing constant of distribution $\pi_l(\boldsymbol{\vartheta})$ so that $\pi_l(\boldsymbol{\vartheta}) = \pi(\boldsymbol{\vartheta})p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_l}/Z_l$. Then one can write

$$m(\boldsymbol{y}) = Z_L = \prod_{l=1}^{L} \frac{Z_l}{Z_{l-1}} \tag{2.27}$$

Equation (2.27) above holds since $Z_0 = \int \pi(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} = 1$. Then, by noticing that

$$\begin{aligned} \int \frac{\pi(\boldsymbol{\vartheta})p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_l}}{\pi(\boldsymbol{\vartheta})p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_{l-1}}}\pi_{l-1}(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta} &= \int \frac{Z_l}{Z_{l-1}}\pi_l(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta} \\ &= \frac{Z_l}{Z_{l-1}} \end{aligned} \tag{2.28}$$

Since at step l the sample $\{\boldsymbol{\vartheta}_t^{(l)}, 1/T\}$ is approximately distributed according to $\pi_l(\boldsymbol{\vartheta}|\boldsymbol{y})$, then a straightforward estimate of (2.28) is given by

$$\frac{\widehat{Z_l}}{Z_{l-1}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi(\boldsymbol{\vartheta}_t^{(l)})p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l)})^{\lambda_l}}{\pi(\boldsymbol{\vartheta}_t^{(l)})p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l)})^{\lambda_{l-1}}} = \frac{1}{T}\sum_{t=1}^{T} p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l)})^{\lambda_l - \lambda_{l-1}}$$

which then yields the following biased estimator of the marginal likelihood

$$\widehat{m}_{SMC}(\boldsymbol{y}) = \prod_{l=1}^{L} \frac{\widehat{Z_l}}{Z_{l-1}} \tag{2.29}$$

Choosing wisely the sequence of temperatures is advised in order to ensure a good variance for the estimate (2.29). Indeed, if the gap between two successive temperatures λ_l and λ_{l+1} is too rough, one can expect the discrepancy between the tempered posteriors π_l and π_{l+1} to cause particle degeneracy. To this end, adaptive strategies are used in for instance Jasra et al. 2011 or Schäfer and Chopin 2013 to derive a suitable jump from temperature λ_l to λ_{l+1}. These methods are based on a readily-available estimate of the Effective Sample Size (ESS) at each iteration l given by

$$ESS(\lambda_l) = \frac{\left(\sum_{t=1}^{T} w_t^{(l)}\right)^2}{\sum_{t=1}^{T} (w_t^{(l)})^2}. \tag{2.30}$$

We here stress the dependence on the next temperature λ_l since

$$w_t^{(l)} = \frac{\pi(\boldsymbol{\vartheta}_t^{(l-1)})p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l-1)})^{\lambda_l}}{\pi(\boldsymbol{\vartheta}_t^{(l-1)})p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l-1)})^{\lambda_{l-1}}} = p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l-1)})^{\lambda_l - \lambda_{l-1}}.$$

Therefore, a good adaptive strategy, as suggested by Buchholz et al. 2021, is to chose λ_l such that $ESS(\lambda_l) = cT$ where c is typically equal to 0.8 so that the effective size of the sample of particles at each step is about 80% that of the initial number of particles. This step can be performed by a simple bisection algorithm, for instance.

Algorithm 6 below give a full implementation of the SMC algorithm described in this section. Note that other tuning parameters can be chosen adaptively, such as the number of times the MCMC kernel is applied to the particles in the *mutation*

step, or the tuning parameters of the mutation kernel itself. Interested readers are referred to Buchholz et al. 2021 for a complete review of such adaptive approaches.

Algorithm 6 : *Adaptive tempered SMC for finite mixtures*

Input : Number of particles T, Markov kernels K_l that are π_l invariant, where $\pi_l(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\vartheta})p(\boldsymbol{y}|\boldsymbol{\vartheta})^{\lambda_l}$
Initialization : $l = 0$, $\lambda_0 = 0$
while $\lambda_l < 1$ **do**
  if $l = 0$ **then**
    for $t = 1, \ldots, T$ **do**
      Sample $\boldsymbol{\vartheta}_t^{(0)} \sim \pi_0$ where π_0 is the prior distribution on $\boldsymbol{\vartheta}$
    end
  end
  else
    for $t = 1, \ldots, T$ **do**
      Mutation. Move particles $\boldsymbol{\vartheta}_t^{(l)} \sim K_l(\tilde{\boldsymbol{\vartheta}}_t^{(l-1)}, d\boldsymbol{\vartheta})$
    end
  end
  Find the next temperature $\lambda_{t+1} > \lambda_t$ adaptively using the ESS given in (2.30) as in Buchholz et al. 2021.
  Reweighting.
  for $t = 1, \ldots, T$ **do**
    $w_t^{(l+1)} = p(\boldsymbol{y}|\boldsymbol{\vartheta}_t^{(l)})^{\lambda_{l+1}-\lambda_l}$
    Set $\vartheta_t^{(l+1)} := \vartheta_t^{(l)}$
  end
  Resampling. Resample with replacement from $\{\boldsymbol{\vartheta}_t^{(l+1)}, w_t^{(l+1)}\}_{i=1}^n$ to obtain a new sample with equal weights $\{\tilde{\boldsymbol{\vartheta}}_t^{(l+1)}, \frac{1}{T}\}$
  $l = l + 1$
end
Output : $\hat{m}_K^{SMC}(\boldsymbol{y}) = \prod_l \frac{1}{T} \sum_{t=1}^T w_t^{(l)}$

2.4 Proposed estimators

In this section, we present two novel estimators of the model evidence for conjugate finite mixture models. One of them is inspired by Chib's method and takes advantage of the partitioning of the data induced by such models. It yields efficient and robust results as well as a great reduction in computational time even for $K > 5$. The second one is an application of Kong et al. 1994's sequential imputation algorithm, which is surprisingly very rarely used as a solution to the problem of evidence estimation for finite mixtures.

2.4.1 Chib's estimator on the partitions (ChibPartitions)

Partitioning comes as a natural by-product of the classical Gibbs sampling algorithm given in Algorithm 1. Indeed, if the output of such an MCMC algorithm is $\{\boldsymbol{\vartheta}^{(t)}, \boldsymbol{z}^{(t)}\}_{t=1}^{T}$, we can denote by $\mathcal{C}(\boldsymbol{z}^{(t)})$ the partition on $[n] = \{1, \dots, n\}$ induced by $\boldsymbol{z}^{(t)} = (z_1^{(t)}, \dots, z_n^{(t)})$, for all $t = 1, \dots, T$. For example, if $n = 4$ and $K = 4$ and for some t, $\boldsymbol{z}^{(t)} = (1, 2, 1, 3)$, then the corresponding partition is $\mathcal{C}(\boldsymbol{z}^{(t)}) = \{\{y_1, y_3\}, \{y_2\}, \{y_4\}, \varnothing\}$, i.e observations y_1 and y_3 are in the same cluster whereas y_2 and y_4 are the only members of their respective cluster. A partition is obviously defined up to a permutation of the labels of the corresponding allocation vector $\boldsymbol{z}^{(t)}$. In particular, note that two different allocation vectors $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$ can yield the same partitioning structure. We note this equivalence relation on the partition space by $\mathcal{C}(\boldsymbol{z}) \doteq \mathcal{C}(\tilde{\boldsymbol{z}})$. For instance $\{\{y_1, y_3\}, \{y_2\}, \{y_4\}, \varnothing\} \doteq \{\{y_2\}, \{y_1, y_3\}, \varnothing, \{y_4\}\}$

The core idea of the ChibPartitions estimator that we propose is to apply the marginal likelihood identity used by Chib's algorithm (2.17) to partitions. The intuition is that partitions are invariant to a permutation of the cluster labels obtained through Gibbs sampling, and hence do not suffer from a potential lack of label switching of the MCMC sampler. This is established in the following proposition.

Proposition 2.1. *For the finite mixture model specified in* (2.5)*, the induced prior distribution on the partition* $\mathcal{C}(\boldsymbol{z})$ *for some allocation vector* $\boldsymbol{z} \in \{1, \dots, K\}^n$ *is*

$$\pi(\mathcal{C}(\boldsymbol{z})) = \frac{K!}{(K - K_+)!} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right) \prod_{k=1}^{K} \Gamma(N_k(\boldsymbol{z}) + \alpha_k)}{\Gamma\left(n + \sum_{k=1}^{K} \alpha_k\right) \prod_{k=1}^{K} \Gamma(\alpha_k)}$$

where $K_+ < K$ *denote the number of unique elements in the vector* $\boldsymbol{z}$*, or alternatively the number of non empty clusters implied by* $\boldsymbol{z}$*.*

Proof. First note that the prior on allocations $\pi(\boldsymbol{z})$ given by (2.13) gives the same weight to all allocation vectors $\boldsymbol{z}$ yielding an equivalent partition. Hence,

$$\pi\left(\mathcal{C}(\boldsymbol{z})\right) = \pi\left(\bigcup_{\tilde{\boldsymbol{z}}:\mathcal{C}(\tilde{\boldsymbol{z}}) \doteq \mathcal{C}(\boldsymbol{z})} \tilde{\boldsymbol{z}}\right) = \sum_{\tilde{\boldsymbol{z}}:\mathcal{C}(\tilde{\boldsymbol{z}}) \doteq \mathcal{C}(\boldsymbol{z})} \pi(\boldsymbol{z}) = \frac{K!}{(K - K_+)!} \pi(\boldsymbol{z})$$

where the last equality comes from $|\{\tilde{\boldsymbol{z}} : \mathcal{C}(\tilde{\boldsymbol{z}}) \doteq \mathcal{C}(\boldsymbol{z})\}| = K!/(K - K_+)!$ □

Note that the partitions induced by $K-$component finite mixtures live in the set $\mathcal{P}_K([n])$, the set of partitions of $\{1,\dots,n\}$ with at most K parts. Now, it is possible to rewrite Chib's identity as

$$m(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{C}^0)\pi(\mathcal{C}^0)}{\pi(\mathcal{C}^0|\boldsymbol{y})} \tag{2.31}$$

for some partition $\mathcal{C}^0 \in \mathcal{P}_K([n])$.

This identity is convenient as under a conditionally conjugate model, the likelihood of a partition is available in closed form and given by

$$p(\boldsymbol{y}|\mathcal{C}(\boldsymbol{z})) = p(\boldsymbol{y}|\boldsymbol{z}) = \prod_{k=1}^{K} \int_{\Theta} \prod_{i:z_i=k} p(y_i|\boldsymbol{\theta}) G_0(d\boldsymbol{\theta}) := \prod_{k=1}^{K} m_k(\boldsymbol{z}) \tag{2.32}$$

with the convention that $m(\varnothing) = 1$. The posterior density of a partition $\mathcal{C}^0$ is unfortunately not available in closed form. However, we can estimate it with the following simple Monte Carlo estimator readily computable from the Gibbs sampler output.

$$\hat{\pi}(\mathcal{C}^0|\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{C}^0 \doteq \mathcal{C}(z^{(t)})\}} \tag{2.33}$$

for some sample $\{\boldsymbol{z}^{(t)}\}$ distributed according to the posterior distribution. This estimator is then plugged back into (2.31) to yield the ChibPartitions estimator $\widehat{m}_{ChibPart}(\boldsymbol{y})$. From a computational viewpoint, comparing two partitions with the equivalence relation $\doteq$ can be done with $\mathcal{O}(n)$ operations.

Possible shortcomings of Algorithm 7 stem from the cardinality of the set $\mathcal{P}_K([n])$ potentially making the estimation of the posterior density difficult. Note that there exists no simple expression giving the exact value of $|\mathcal{P}_K([n])|$ for all K and n but it can be written as

$$|\mathcal{P}_K([n])| = \sum_{k=1}^{K} S(n,k) := \sum_{k=1}^{K} \frac{1}{k!}\sum_{l=0}^{k} (-1)^l \binom{k}{l} (k-l)^n$$

where $S(n,k)$ are the Stirling numbers of the second kind, counting the number of ways to partition a set of n distinguishable objects into k nonempty subsets (see, e.g., Graham et al. 1989). This number is increasing very quickly with n and k. Indeed, $|\mathcal{P}_K([n])| \approx n^{k-1}/[(k-1)!k!]$, if $n >> k$. It is therefore crucial to choose $\mathcal{C}^0$ to be the estimated MAP or a similar high-posterior probability partition to achieve a reduced variance for the estimator (2.33). As supported by the simulations in Section 2.5, such a strategy appears to be sufficient to ensure a robust estimator of the posterior density. One might fear that the connection of $\mathcal{C}^0$ with the MCMC sample $\{C(z^{(t)})\}_t$ could introduce a bias in (2.33). Simulations in Section 2.5 give no indication of such a phenomenon. Were a bias detected, then a simple strategy would be to choose $\mathcal{C}^0$ within an independent MCMC sample simulated in parallel to the one used to compute (2.33).

Compared with alternative corrections of Chib's 1995 method, such as the fully permuted Chib's estimator described earlier (hereafter ChibPerm) which requires

$\mathcal{O}(TK!)$ likelihood evaluations, ChibPartitions only needs $\mathcal{O}(T)$, where T is the length of the Markov chain in Algorithm 7.

Proposition 2.2. *An estimate of the variance of* $\log \widehat{m}_{ChibPart}(\boldsymbol{y})$ *is given by*

$$\widehat{\mathbb{V}}(\log \widehat{m}_{ChibPart}(\boldsymbol{y})) = (\hat{\pi}(\mathcal{C}^0|\boldsymbol{y}))^{-2}\widehat{\mathbb{V}}(\hat{\pi}(\mathcal{C}^0|\boldsymbol{y}))$$

where

$$\widehat{\mathbb{V}}(\hat{\pi}_K(\mathcal{C}^0|\boldsymbol{y})) = \frac{1}{T}\left(\sigma_0 + 2\sum_{s=1}^{q}\left(1 - \frac{s}{q+1}\right)\sigma_s\right)$$

for $\sigma_s = \frac{1}{T}\sum_{t=s+1}^{T}\left(\mathbb{1}_{\{\mathcal{C}^0 \doteq \mathcal{C}(\boldsymbol{z}^{(t)})\}} - \hat{\pi}_K(\mathcal{C}^0|\boldsymbol{y})\right)^2$ *and* q *large enough.*

Proof. The expression of the auto-correlation consistent variance estimator $\widehat{\mathbb{V}}(\hat{\pi}_K(\mathcal{C}^0|\boldsymbol{y}))$ is an immediate application of Newey and West 1986. Then, the estimate of $\log \widehat{m}_{ChibPart}(\boldsymbol{y})$ is obtained by the Delta method. □

Algorithm 7 : *ChibPartitions estimator for conditionally conjugate mixtures*

Input : $(\boldsymbol{z}^{(t)})_{t=1}^{T}$ from an MCMC targeting $\pi(\boldsymbol{z}|\boldsymbol{y})$
for $t = 1, \ldots, T$ **do**
  Compute $\tilde{\pi}(\mathcal{C}(\boldsymbol{z}^{(t)})|\boldsymbol{y}) = p(y|\mathcal{C}(\boldsymbol{z}^{(t)}))\pi(\mathcal{C}(\boldsymbol{z}^{(t)}))$ using Proposition 2.1 and (2.32)
  Set $\mathcal{C}^0 = \underset{t=1,\ldots,T}{\operatorname{argmax}} \{\tilde{\pi}(\mathcal{C}(\boldsymbol{z}^{(t)})|\boldsymbol{y})\}$
end
Compute $\hat{\pi}(\mathcal{C}^0|\boldsymbol{y}) = (1/T)\sum_{t=1}^{T}\mathbb{1}(\{\mathcal{C}^0 \doteq \mathcal{C}(z^{(t)})\})$
Output : $\widehat{m}_{ChibPart}(\boldsymbol{y}) = p(\boldsymbol{y}|\mathcal{C}^0)\pi(\mathcal{C}^0)/\hat{\pi}(\mathcal{C}^0|\boldsymbol{y})$

2.4.2 Sequential Importance Sampling

The Sequential Importance Sampling (SIS) algorithm described here can be included in the SMC framwork described in Section 2.3.4. It stems from Kong et al. 1994 that addresses the issue of missing data problems by sequential imputation, using a latent variable $\boldsymbol{z}$, representing the missing part of the data. The authors show that for a particular choice of importance distribution π^*, a SIS procedure yields a direct estimator of the evidence $m(\boldsymbol{y})$. This is applicable whenever one can sample easily from distributions $p(z_i|\boldsymbol{y}_{1:i}, \boldsymbol{z}_{1:i-1})$ for all $i \geq 2$ where $\boldsymbol{z}_{1:i} = (z_1, \ldots, z_i)$ and whenever the prequential predictive densities $p(y_i|\boldsymbol{y}_{1:i-1}, \boldsymbol{z}_{1:i-1})$ are available in closed form for all $i \geq 2$. The link to the latent cluster membership of finite mixture models $\boldsymbol{z}$ is immediate and shall be highlighted later in this Section.

The core idea is to define $\pi^* := \pi^*(z_1, \ldots, z_n|\boldsymbol{y}) = p(z_1|y_1)\prod_{i=2}^{n} p(z_i|\boldsymbol{y}_{1:i}, \boldsymbol{z}_{1:i-1})$ as an approximation to the posterior distribution in which the latent variable $\boldsymbol{z}$ is

imputed sequentially. Then one can notice that

$$\begin{aligned}\pi(\boldsymbol{z}|\boldsymbol{y}) \times \frac{1}{\pi^*(\boldsymbol{z}|\boldsymbol{y})} &= \frac{p(\boldsymbol{z},\boldsymbol{y})}{m(\boldsymbol{y})} \times \left(\frac{p(y_1)}{p(z_1,y_1)} \frac{p(z_1,y_1,y_2)}{p(z_1,y_1,z_2,y_2)} \cdots \frac{p(\boldsymbol{z}_{1:n-1},\boldsymbol{y})}{p(\boldsymbol{z},\boldsymbol{y})} \right) \\ &= \frac{p(y_1)\prod_{i=2}^n p(y_i|\boldsymbol{z}_{1:i-1}\boldsymbol{y}_{1:i-1})}{m(\boldsymbol{y})} \\ &= \frac{w(\boldsymbol{z},\boldsymbol{y})}{m(\boldsymbol{y})}\end{aligned}$$

where $w(\boldsymbol{z},\boldsymbol{y}) := p(y_1)\prod_{i=2}^n p(y_i|\boldsymbol{z}_{1:i-1},\boldsymbol{y}_{1:i-1})$.
This leads to the following identity

$$\begin{aligned}\int w(\boldsymbol{z},\boldsymbol{y})\pi^*(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z} &= \int m(\boldsymbol{y})\pi(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z} \\ &= m(\boldsymbol{y})\end{aligned}$$

which implies the unbiased estimator of the marginal likelihood of the data $\boldsymbol{y}$

$$\widehat{m}_{SIS}(\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^T w(\boldsymbol{z}^{(t)},\boldsymbol{y})$$

for a sample $\{z^{(t)}\}_{t=1}^T$ from $\pi^*(\boldsymbol{z}|\boldsymbol{y})$.

The strength of this estimator is that, just like ChibPartitions, it does not suffer from label switching or the lack threreof.

Proposition 2.3. *An estimate of the standard deviation of* $\log \widehat{m}_{SIS}(\boldsymbol{y})$ *is given by*

$$\widehat{sd}(\log \widehat{m}_{SIS}(\boldsymbol{y})) = \frac{1}{\sqrt{T}} \times \frac{\widehat{sd}(\boldsymbol{w})}{\widehat{m}_{SIS}(\boldsymbol{y})}$$

Proof. The proof is an immediate application of the Delta method and can be found in Irwin et al. 1994. □

As mentioned earlier, it is immediate to apply the SIS framework to the conditionnaly conjugate finite mixture model, as also noticed by Carvalho et al. 2010. The only requirement is to be able to compute π^* for the latent cluster membership variable $\boldsymbol{z}$, which we give below.

Proposition 2.4. *In the conditionnaly conjugate mixture model specified in* (2.5)*, for all* $1 < i \leq n$*,*

$$p(y_i|\boldsymbol{z}_{1:i-1},\boldsymbol{y}_{1:i-1}) = \sum_{k=1}^K \frac{m_k(\boldsymbol{z}_{1:i-1}\cup\{z_i=k\})}{m_k(\boldsymbol{z}_{1:i-1})} \frac{N_k(\boldsymbol{z}_{1:i-1})+\alpha_k}{i-1+\sum_{k=1}^K \alpha_k} \tag{2.34}$$

and for all $k = 1,\ldots,K$*,*

$$p(z_i = k|\boldsymbol{y}_{1:i},\boldsymbol{z}_{1:i-1}) \propto \frac{m_k(\boldsymbol{z}_{1:i-1}\cup\{z_i=k\})}{m_k(\boldsymbol{z}_{1:i-1})} \frac{N_k(\boldsymbol{z}_{1:i-1})+\alpha_k}{i-1+\sum_{k=1}^K \alpha_k} \tag{2.35}$$

Proof. For all $1 < i \leq n$,

$$\begin{aligned}
p(y_i|\boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1}) &= \int\int p(y_i|\boldsymbol{\varpi}, \boldsymbol{\theta}) \prod_{k=1}^{K} \Pi(d\theta_k|\boldsymbol{y}_{1:i-1}, \boldsymbol{z}_{1:i-1}) \Pi(d\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}, \boldsymbol{\alpha}) \\
&= \sum_{k=1}^{K} \int\int \varpi_k f(y_i|\theta_k) \pi(\theta_k|\boldsymbol{y}_{1:i-1}, \boldsymbol{z}_{1:i-1}) d\boldsymbol{\theta} \Pi(d\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}, \boldsymbol{\alpha}) \\
&= \sum_{k=1}^{K} \frac{m_k(\boldsymbol{z}_{1:i-1} \cup \{z_i = k\})}{m_k(\boldsymbol{z}_{1:i-1}))} \int \varpi_k \Pi(d\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}, \alpha) \\
&= \sum_{k=1}^{K} \frac{m_k(\boldsymbol{z}_{1:i-1} \cup \{z_i = k\})}{m_k(\boldsymbol{z}_{1:i-1}))} \frac{N_k(\boldsymbol{z}_{1:i-1}) + \alpha_k}{i - 1 + \sum_{k=1}^{K} \alpha_k}
\end{aligned}$$

where the second equality comes directly from equations (2.9) and (2.12), while the integral of the third equality is simply the expectation of the Dirichlet posterior of the weights given in (2.8). Finally, (2.35) is proved by noticing that

$$\begin{aligned}
p(z_i = k|\boldsymbol{y}_{1:i}, \boldsymbol{z}_{1:i-1}) &\propto p(z_i = k, y_i|\boldsymbol{y}_{1:i-1}, \boldsymbol{z}_{1:i-1}) \\
&= \int\int p(y_i|\theta_k, z_i = k) \pi(z_i = k|\boldsymbol{\varpi}) \pi(\theta_k|\boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1}) \\
&\quad \times \pi(\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}) d\theta_k d\boldsymbol{\varpi} \\
&= \int\int \varpi_k f(y_i|\theta_k) \pi(\theta_k|\boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1}) \pi(\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}) d\theta_k d\boldsymbol{\varpi} \\
&= \frac{m_k(\boldsymbol{z}_{1:i-1} \cup \{z_i = k\})}{m_k(\boldsymbol{z}_{1:i-1}))} \int \varpi_k \Pi(d\boldsymbol{\varpi}|\boldsymbol{z}_{1:i-1}, \alpha) \\
&= \frac{m_k(\boldsymbol{z}_{1:i-1} \cup \{z_i = k\})}{m_k(\boldsymbol{z}_{1:i-1}))} \frac{N_k(\boldsymbol{z}_{1:i-1}) + \alpha_k}{i - 1 + \sum_{k=1}^{K} \alpha_k}
\end{aligned}$$

□

Despite the efficiency of such samplers that will be highlighted in the following section, note that SIS, and more generally SMC approaches, are not as popular as Chib's algorithm or Bridge sampling.

We give in Algorithm 8 an implementation of SIS for a conditionally conjugate mixture model.

Algorithm 8 : *SIS for the conditionally conjugate mixture model*

Input : Number of iterations T
for *t=1,...,T* **do**
    /* Initialization */
    Sample $z_1^{(t)}$ from $\pi(z_1|y_1)$
    Compute $p(y_1) = m(\{y_1\})$, set $w^{(t)} \leftarrow p(y_1)$
    for *i=2,...,n* **do**
        Sample $z_i^{(t)}$ using (2.35)
        Set $w^{(t)} \leftarrow w^{(t)} p(y_i|\boldsymbol{z}_{1:i-1}^{(t)}, \boldsymbol{y}_{1:i-1})$ using (2.34)
    end
end
Return $\widehat{m}_{SIS}(\boldsymbol{y}) = 1/T \sum_{t=1}^{T} w^{(t)}$

2.5 Simulation study

In this section, we assess and compare our proposed estimators to Chib's algorithm, Bridge sampling and SMC. To do so, we consider different experimental designs, involving large values of K and n.

Unless specified otherwise, we fit Normal mixture models and use the *conditionally conjugate* Normal-Inverse Gamma prior for the location and scale parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ as described in Example 1 and equation (2.10) that is defined for all $k = 1, \ldots, K$ by $\sigma_k^2 \sim \Gamma^{-1}(a, b)$ and $\mu_k|\sigma_k^2 \sim \mathcal{N}(\mu_0, \sigma_k^2/\lambda)$ where Γ^{-1} is the inverse gamma distribution in the shape and scale parametrization. The hyperparameters (a, b, μ, λ) are derived empirically following recommendations from Raftery 1996 : $a = 1.28, b = 0.36(\overline{y^2} - \overline{y}^2), \mu_0 = \bar{y}, 1/\lambda = (y_{max} - y_{min})/2.6$. Finally the prior on the mixture weights is chosen to be Dirichlet with concentration parameter $\boldsymbol{\alpha} = \mathbf{1}_K$. Note that this choice of prior ensures that $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z}) := \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2|\boldsymbol{y}, \boldsymbol{z})$ is available in closed form, which is a prerequisite to most of the algorithms we wish to implement.

2.5.1 Experiment 1 : Galaxies data

The first experiment we design aims at assessing the relative performance of our suggested ChibPartitions and SIS algorithms in a basic setting. As is usually done in the mixture modeling community, we use the benchmark `galaxies` data set that contains the radial velocity of 82 galaxies.

Figure 2.6 shows boxplots of the estimates of the marginal likelihood given by the above methods for an increasing number of mixture components K. In each scenario, the simple arithmetic mean estimator, computed with a very large number of simulations $\{\boldsymbol{\vartheta}^{(t)}\}_{t=1}^T$ from the prior $\pi(\boldsymbol{\vartheta})$, is defined as $\widehat{m}_{AME}(\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^T p(\boldsymbol{y}|\boldsymbol{\vartheta}^{(t)})$. This prohibitively time-consuming estimator is here solely for a benchmarking purpose. Except for this estimator, all other algorithms are allocated as much time as required for them to converge, provided this time is reasonable. For instance, bridge sampling and the fully-permuted Chib estimator (ChibPerm) are not included for $K = 6$ and $K = 8$ as they fail to converge in a time comparable to the other methods. Hence, Figure 2.6 only provides insights about which methods provide reliable estimates for an increasing number of mixture components. For $K = 3$, although most estimates agree on a common value for $m(\boldsymbol{y})$, Chib's method is almost exactly off by a factor $\log 3! \approx 1.79$, which is the consequence of an almost complete lack of label switching during the Gibbs sampling step, as discussed earlier. The sum over all permutations produced by the fully-permuted Chib's estimator (ChibPerm) makes up for this bias, and so does ChibRandPerm, which sums over 100 randomly sampled permutations. As K grows, all classical methods except adaptive SMC fail to estimate the marginal likelihood while our candidates ChibPartitions and SIS are consistently pointing to the reference value given by the arithmetic mean estimator. For $K = 5$, the time (in seconds) taken by each of the four successful algorithms (ChibPartitions, SIS, adaptive SMC and Bridge sampling) to yield one estimate of the marginal likelihood as displayed in Figure 2.6 is respectively 425, 472, 7539, 57983. The time given corresponds to the average computational time over 20 repetitions of each estimator on one core Intel(R) Xeon(R) CPU E5-2630

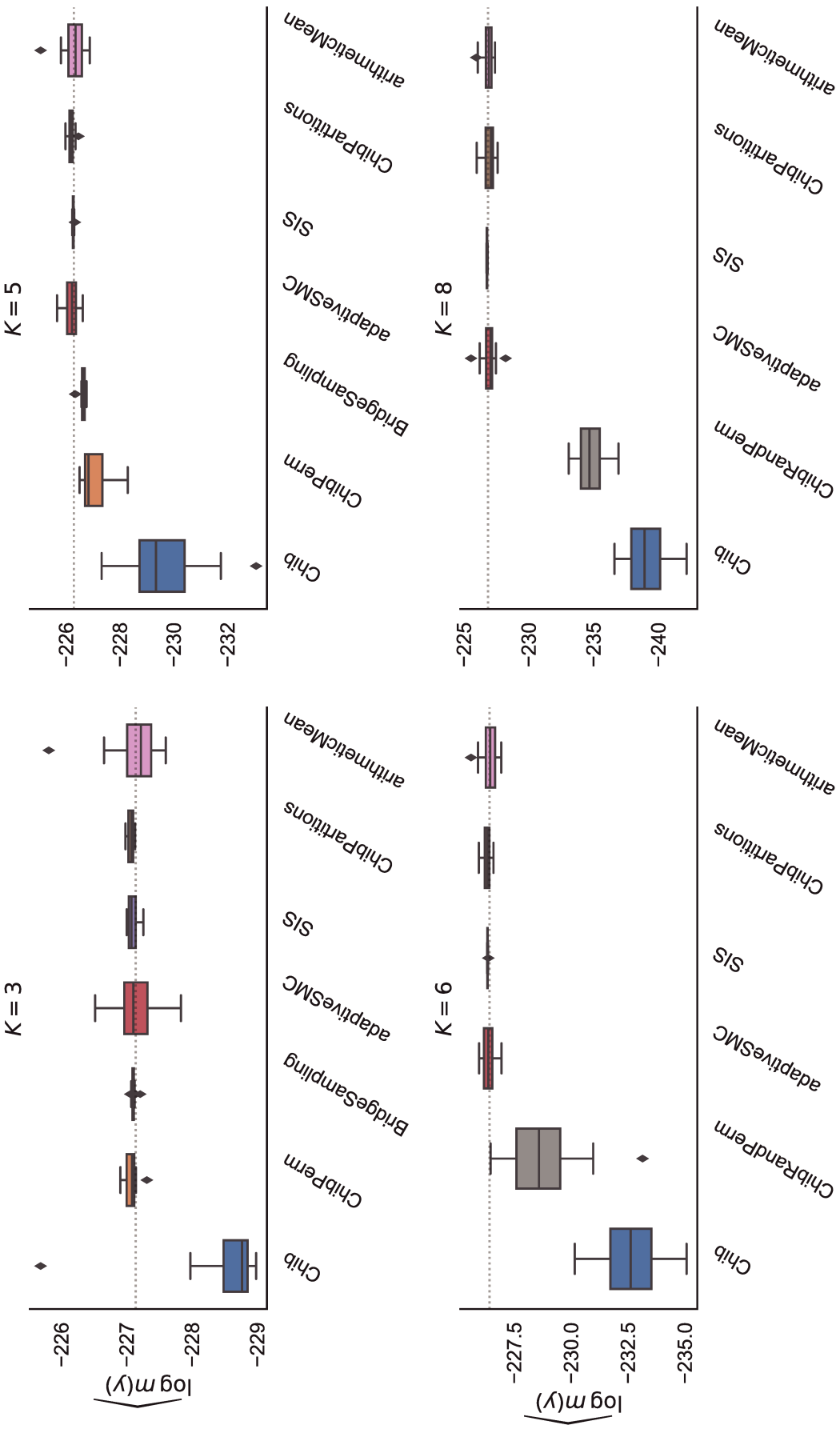


Figure 2.6: Boxplot of the marginal likelihood estimators 20 repetitions each. Dashed line : mean of the Arithmetic Mean estimator.

v4 @ 2.20GHz. As is expected theoretically, one can observe a massive decrease in computational time offered by ChibPartitions and SIS with respect to Bridge Sampling and the fully permuted Chib's estimator. Although it is not done in this experiment, note that it is straightforward to parallelize the embarrassingly parallel SIS algorithm and thus to further reduce its computational time. Figure 2.8 shows the evolution of the Mean Squared Error (MSE) as a function of time for the 5 component-mixture model on the `galaxies` data. Note that methods SIS and SMC are not implemented in their parallelized version. Despite this, SIS clearly outperforms the other algorithms, closely followed by ChibPartitions.

Figure 2.7 gives boxplots of the marginal likelihoods estimators given by SIS for different values of K and indicates that a mixture of 5 components is best supported by the `galaxies` data, for our choice of prior.

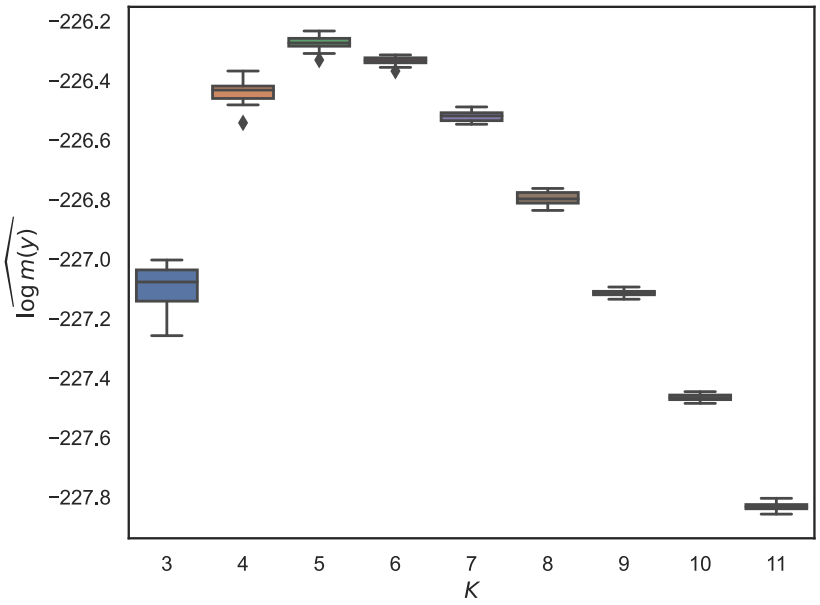


Figure 2.7: `galaxies` data. Boxplots of the SIS marginal likelihood estimators for different values of K. 20 repetitions each.

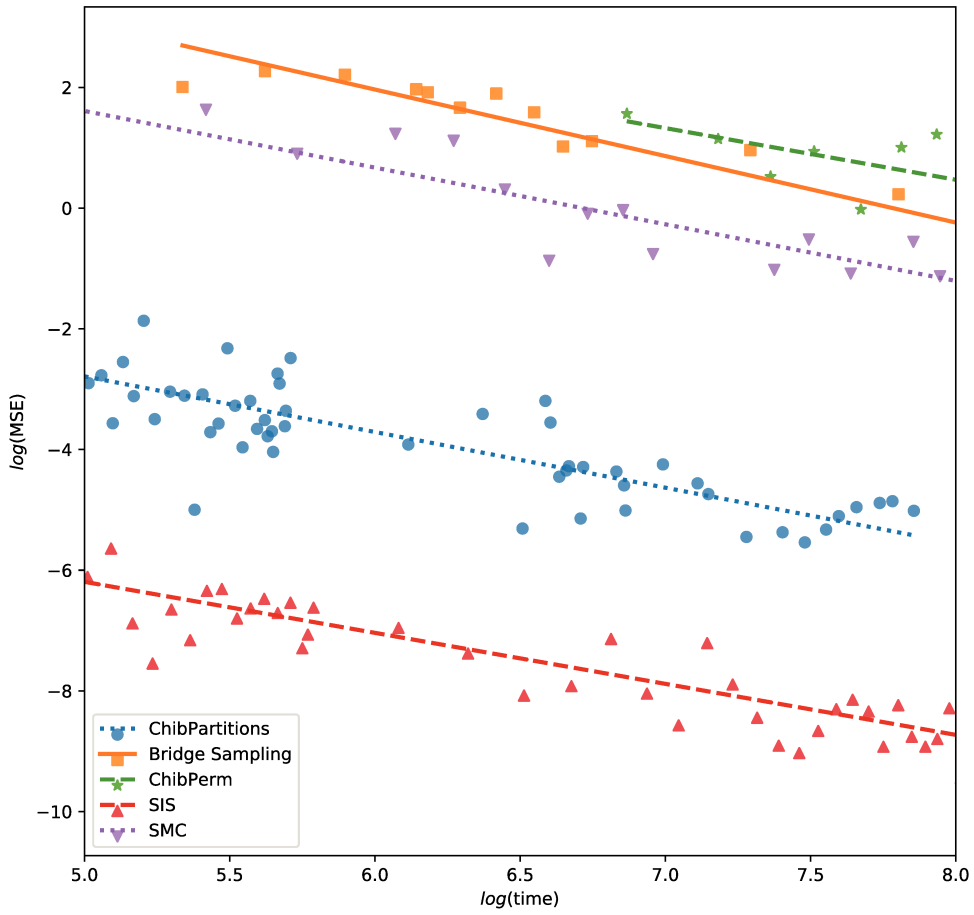


Figure 2.8: `galaxies` data. MSE vs time. The dashed lines represent the Ordinary Least Square line fitting the log of the MSE and the log of the time for each of the different methods.

2.5.2 Experiment 2 : Synthetic data, $n = 1000$ and $n = 2000$

For more realistic applications, our goal is to find which algorithms scale well as both K and n get large. To our knowledge, no earlier work has been conducted towards identifying reliable estimators for this kind of challenging scenarios.

1000 observations. We generate a data set $\boldsymbol{y}$ of size 1000 from the following 3-component Normal mixture

$$0.3\mathcal{N}(0,1) + 0.3\mathcal{N}(3,1) + 0.4\mathcal{N}(6,1).$$

Normal mixtures are then fitted onto the data with $K = 3, 7, 10$ and 13 components. Given the results of the previous experiment in which all but three methods failed

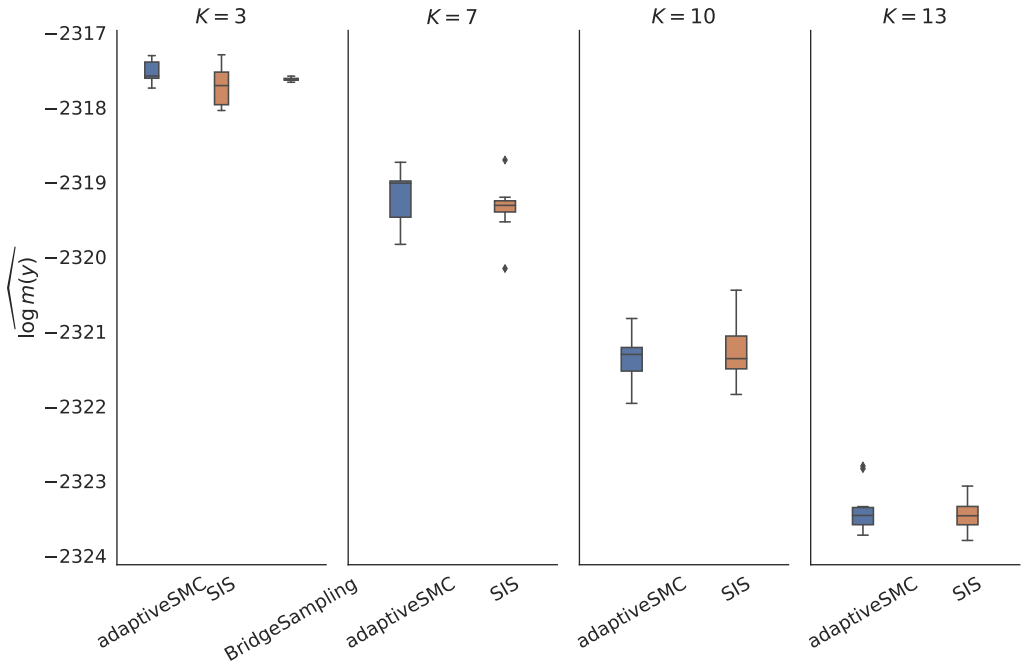


Figure 2.9: Experiment 2, $n = 1000$. Boxplots with 20 replications for each method considered

to converge in a reasonable time, we only consider ChibPartitions, SIS, SMC and Bridge sampling for this more difficult scenario. It is difficult to compare these four methods on an equal footing since there is no one-to-one connection between their respective hyperparameters. The results obtained on Figures 2.9 and 2.10 are roughly the best possible outcome for each algorithm. That is, when increasing the value of a hyperparameter, thus allocating more computational time, we do not observe a significant improvement for the observed Monte Carlo variance.

The obtained results are presented on Figures 2.9 and 2.10. ChibPartitions is presented in a separate plot for readability reasons. We can indeed see that it suffers from a pathological variance, probably due to the high cardinality of the set of partitions $\mathcal{P}_3([1000])$. This in turn leads to a strong downward bias on the log scale. The other methods considered all seem to agree on a common value for the marginal likelihood for all considered values of K. Note that the Bridge sampling estimator is not included in the scenarios where $K > 3$ as it failed to give an output in a reasonable time. For $K = 3$ however, Bridge Sampling seems to be giving the most accurate estimate of $\log \widehat{m}(\boldsymbol{y})$. The average computational times to obtain one marginal likelihood estimate is given in Table 2.1. We see that the computational time of SIS is much lower than that of adaptive SMC, for a comparable estimation precision.

2000 observations. The data generating process for this scenario is a 6-component mixture of Normal distributions with means $\boldsymbol{\mu} = (2.51, -6.22, -5.28, -4.54, 2.75, 11.46)$, weights $\boldsymbol{\varpi} = (0.20, 0.01, 0.27, 0.20, 0.18, 0.14)$ and scales $\boldsymbol{\sigma} = (2, 2, 2, 2, 2, 2)$. Given the previous experiment, we do not include ChibPartitions in our simulations

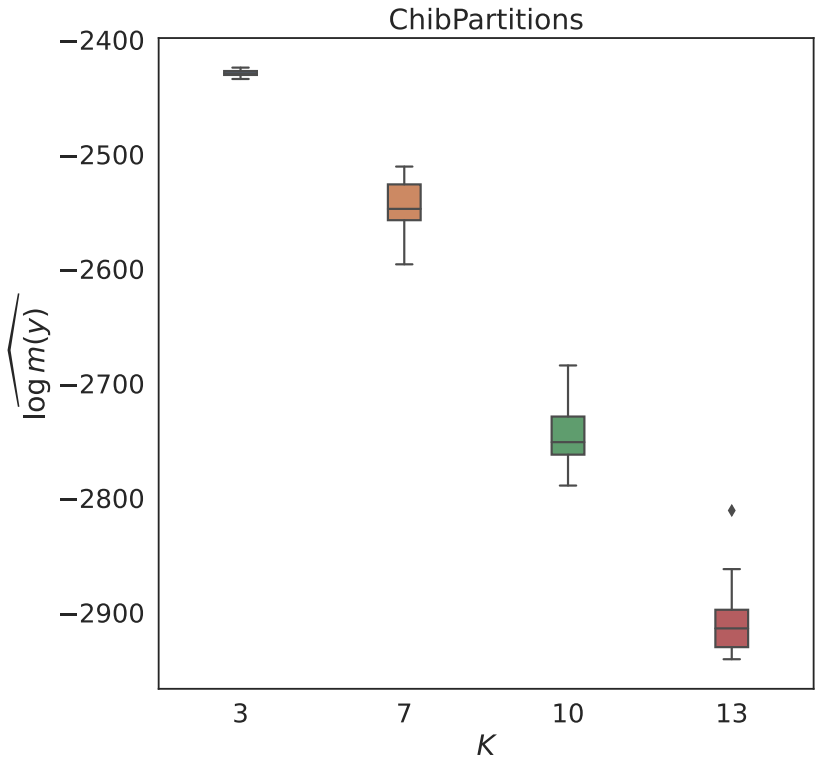


Figure 2.10: Experiment 2 $n = 1000$. Boxplots of ChibPartitions with 20 replications

Algorithm K	Time in seconds ChibPartitions	SIS	SMC	Bridge Sampling
3	110	230	3565	1008
7	324	250	7870	-
10	532	434	11907	-
13	778	674	17496	-

Table 2.1: Experiment 2. Average time needed to obtain one of the marginal likelihood estimates used in the boxplots of Figures 2.9 and 2.10.

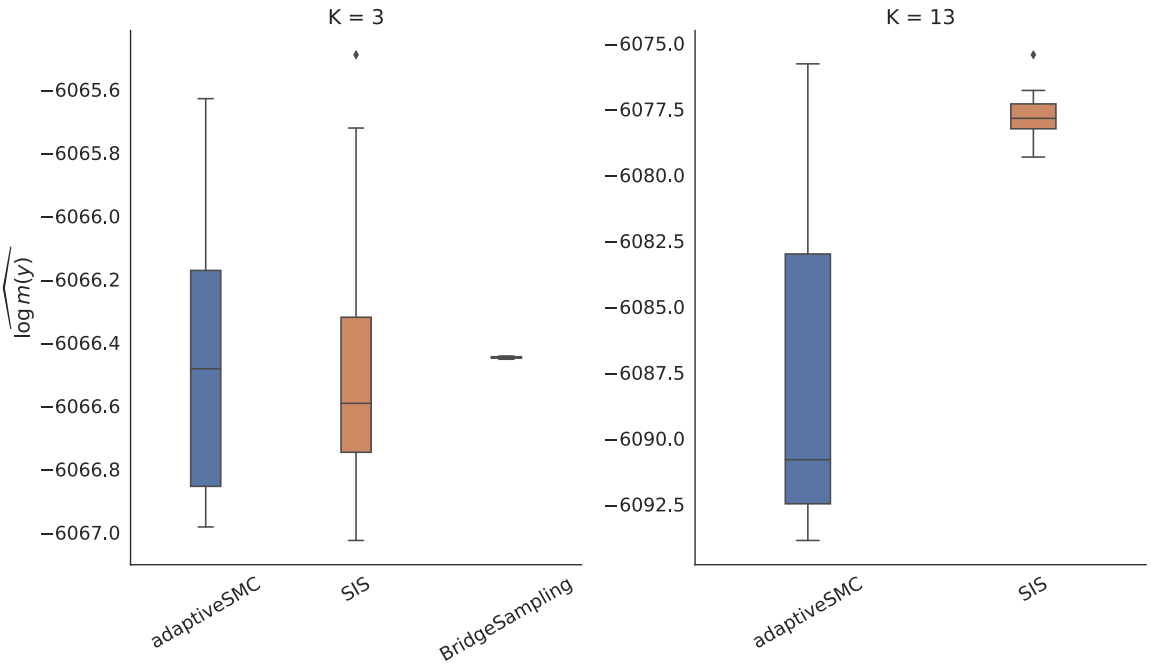


Figure 2.11: Experiment 2 $n = 2000$. Boxplots with 20 replications for each method considered

	Time in seconds		
Algorithm	SIS	SMC	Bridge Sampling
K			
3	153	5890	2870
13	903	34468	-

Table 2.2: Experiment 2. Average time needed to obtain one of the marginal likelihood estimates used in the boxplots of Figure 2.11.

and focus on SMC, SIS and Bridge Sampling. The latter, as shown on 2.11, has a considerably small variance for the case $K = 3$, but is unfortunately unable to converge in a reasonable time for $K = 13$. Although they show comparable results for $K = 3$, SIS clearly outperforms SMC both in terms of variance and computational time (cf Table 2.2) for the most complex scenario where $K = 13$.

2.5.3 Experiment 3 : Synthetic data, $n = 1000$, well-specified mixture

In this experiment, we try to assess the performance of our methods on a well-specified six-component mixture of normal distributions. The Data Generating Process (DGP) is the six-component normal mixture with equal weights, means $\boldsymbol{\mu} = (0, 6, 12, 18, 24, 30)$ and variances $\boldsymbol{\sigma}^2 = (1, 1, 1, 1, 1, 1)$ from which we generate $n = 1000$ data points $\boldsymbol{y}$. On the other hand, we fit a conditionally conjugate normal mixture model with $K = 6$ and hyperparameters as the ones chosen in Experiment

1 except for $b = 3$, so that the $\mathbb{E}(\sigma_k^2) = b/(a-1)$ is not too large a priori. Figure 2.12 below gives the histogram of the generated data.

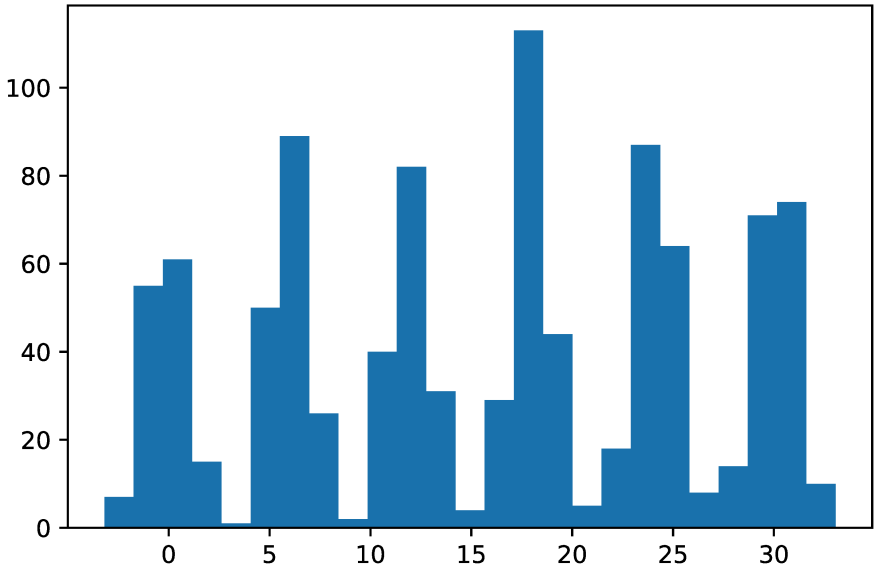


Figure 2.12: Histogram of the generated data for Experiment 3

As can be seen, such a DGP yields rather well-separated data. The intuition is that in this setting, there exist only a few partitions that are likely a posteriori, which should yield great performances of ChibPartitions, despite the dimension of the latent space of partitions which is about $1000^5/(5!6!)$.

For this experiment, we do not assess the performances of algorithms ChibPerm or Bridge Sampling as their computational cost is too high given the number of mixture components $K = 6$. The results are given in Figure 2.13 where we see 20 repetitions of algorithms ChibPartitions, SIS and SMC. We clearly see that SMC suffers from a rather large variance while ChibPartitions is pointing very accurately to a value of about -3246.5. It is worth noting that the computational times required to obtain one such estimate are respectively 174s, 8067s, and 22401s for ChibPartitions, SIS, and SMC.

These computational times can be further investigated on Figure 2.14 where we see that the Mean Squared error of ChibPartitions decreases much faster as a function of time compared to SIS or SMC.

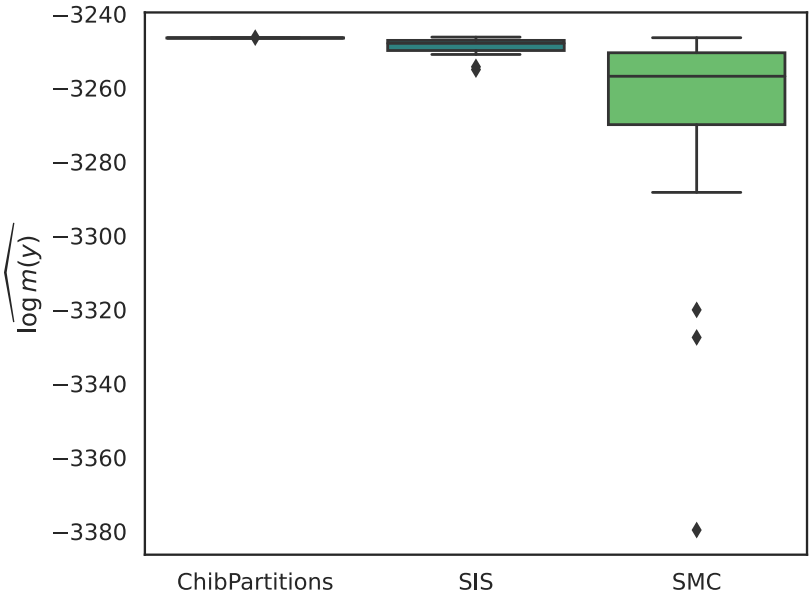


Figure 2.13: Boxplots for the Experiment 3. 20 repetitions for each algorithm.

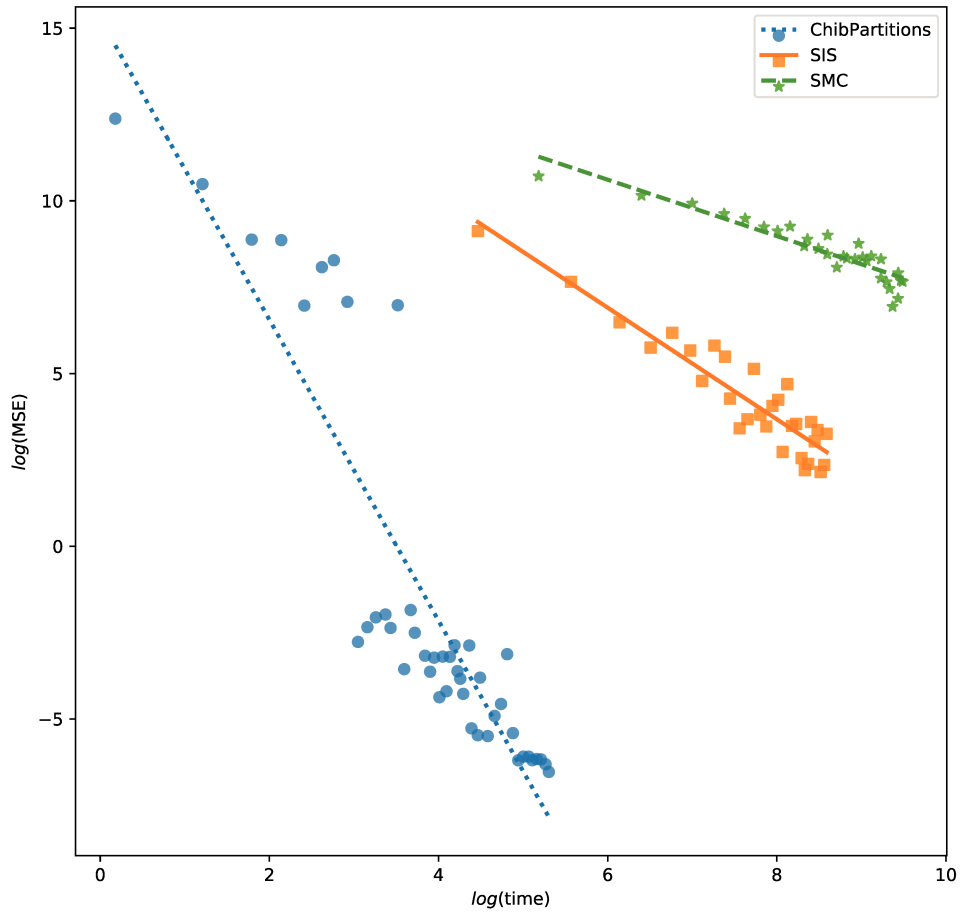


Figure 2.14: Experiment 3. MSE vs time. The dashed lines represent the Ordinary Least Square line fitting fitting the log of the MSE and the log of the time for each of the different methods. The reference value to compute the MSE is the value of ChibPartitions with 10^5 iterations.

2.5.4 Experiment 4 : Bayes Factor convergence

In this final experiment, we illustrate how computing the Bayes Factor to find the true number of components K_0 of some data arising from a K_0-mixture of Normal distributions is a consistent procedure. This is in fact a well-known fact and interested readers can refer to Chib and Kuffner 2016 for instance.
To do so, we generate 100 data sets of size 1000 from the 3-component mixture

$$P_{K_0} := 0.3\mathcal{N}(0,1) + 0.3\mathcal{N}(2,1) + 0.4\mathcal{N}(4,1)$$

and, for increasing values of n, fit two competing Normal mixture models with respectively $K_0 = 3$ and $K \neq 3$ components. The Bayes Factor is then computed

using SIS and the resulting so-called *Bayes Factor paths* are displayed on Figure 2.15 where we have considered the cases $K = 2, 5, 7$ and 10. For all scenarios but $K = 5$, the procedure seems consistent since it seems that $P_{K_0}(\log BF > 0) \longrightarrow 1$. While it is less straigthforward to make such a conclusion for the case $K = 5$, we clearly see that the bulk of the log Bayes Factor trajectories seems to head towards positive values as n increases.

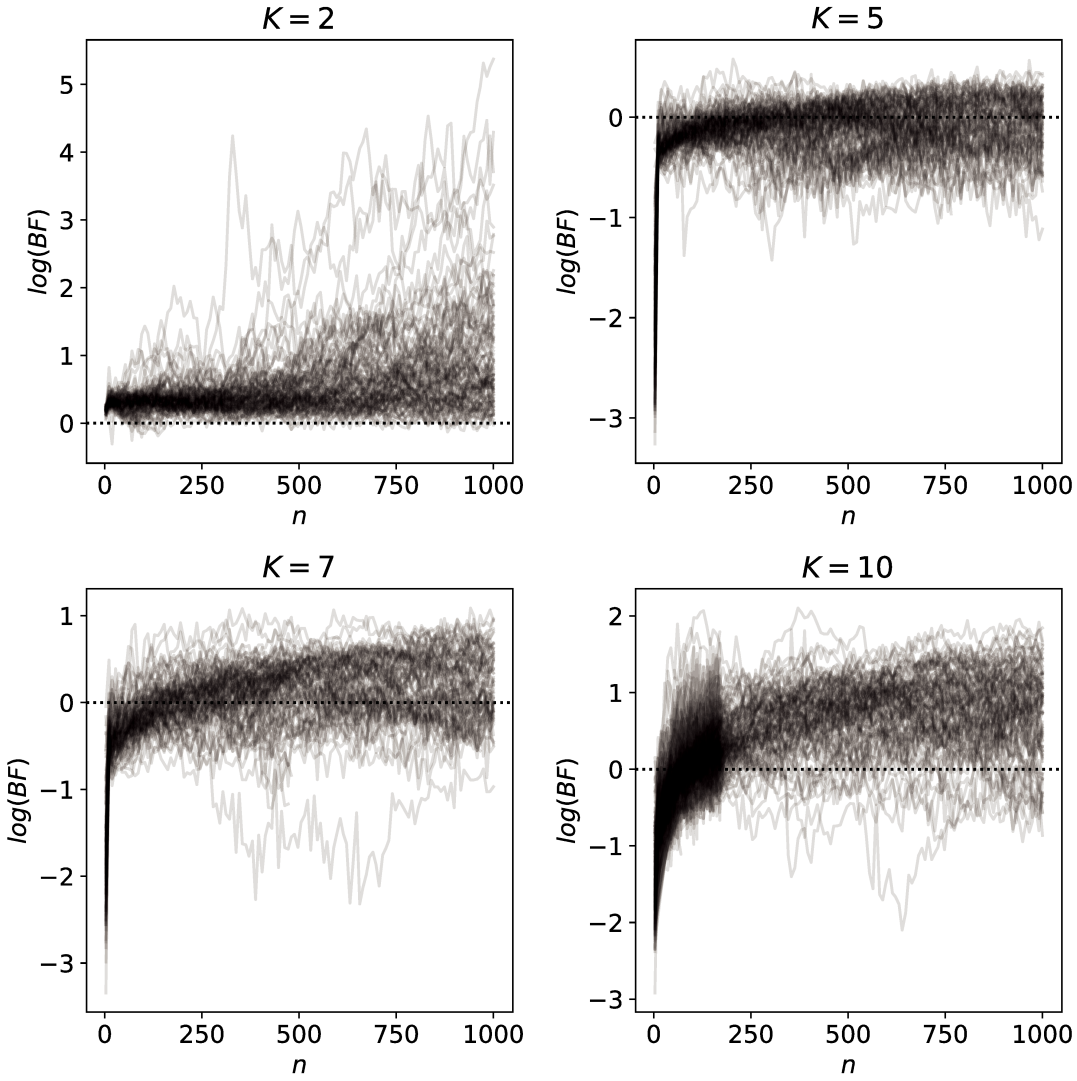


Figure 2.15: Experiment 4. Bayes Factors paths for 100 data sets from P_{K_0} of a K_0-component mixture against a $K-$component mixture.

2.6 Conclusion and perspectives

In this chapter, we have explained why estimating the marginal likelihood of $K-$component finite mixture models is a complex task that generally requires *ad hoc* solutions. Classical estimators usually suffer from a sort of curse of dimensionality phenomenon as their complexity typically grows as $K!$, due to the multimodality of the posterior distribution.

We identified two methods, namely ChibPartitions and SIS, that scale with the number of mixture components and/or with the number of observations, as opposed to classical algorithms. We have investigated several scenarios to better understand the strengths and flaws of these algorithms. In particular, it seems that ChibPartitions is more efficient than any other method provided the considered mixture model is well-specified. On the other hand, SIS outperforms the other algorithms in more ill-specified settings and provides reliable estimates for all kind of scenarios. Furthermore, the adaptive SMC algorithm that we have presented shows rather good performance and can be used in the non-conjugate case.

An interesting research avenue would consist in deriving a more elaborate estimate for the posterior on the partitions $\pi(\mathcal{C}_0|\boldsymbol{y})$. This could indeed greatly improve the performance of the ChibPartitions algorithm.

Chapter 3

Evidence asymptotics and estimation for infinite mixtures

Abstract

In this chapter, we investigate the frequentist properties of the marginal likelihood of a Dirichlet Process mixture, and establish the consistency of the Bayes factor when comparing a parametric family of finite mixtures against the nonparametric 'strongly identifiable' Dirichlet Process Mixture model. We then consider the problem of estimating numerically the marginal likelihood of the Dirichlet Process mixture (DPM) model, a.k.a the *infinite* mixture model. This is a difficult problem that is still lacking a fully satisfactory resolution. In particular, we evaluate the algorithm proposed by Basu and Chib 2003 and suggest an alternative based on an idea of Geyer 1994.

Contents

3.1 Introduction

The Dirichlet Process Mixture model (DPM), or 'infinite' mixture model, first introduced by Ferguson 1983, has become one of the main tools in the field of Bayesian nonparametrics. Its range of applications is broad, from multivariate clustering (see, e.g., Crépet and Tressou 2011) to Bayesian density estimation (Neal 1992, Rabaoui et al. 2012). Such models are of particular interest when practitioners are not willing to make strong assumptions on the underlying *a priori* distribution of the mixture parameters. In this setting, their prior distribution is assumed to be a realization of the Dirichlet Process (DP), formalized by Ferguson 1973, which can be viewed as a prior on the space of discrete distributions.

Bayesian model selection is primarily done by comparing the *marginal likelihoods* of the data (a.k.a *model evidence*) for competing models, defined as

$$m(\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}|\theta)\Pi(d\theta)$$

for some data $\boldsymbol{y}$, a density function f, prior Π, and parameter space Θ. In the context of mixture modeling, being able to compute the evidence of a model is of great importance and has several applications of interest. For instance, the usual problem of statistical modeling of determining whether an $n-$sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ arises from a particular parametric family of distributions can be formalized as a goodness of fit test against a nonparametric alternative, like the DPM. For example, Tokdar and Martin 2021 designs a particular DPM alternative for testing normality. Other applications include Argiento et al. 2010 who attempts to find the best-fitting mixing measure by minimizing the Bayes Factor for parametric against nonparametric alternatives in the context of Bayesian density estimation, or Ray and Mallick 2006 who compute the marginal likelihood of a DPM in order to choose a partition of the data. The Bayes Factor was proven to be consistent for a point null hypothesis against a large class of nonparametric alternatives by Dass and Lee 2004. Mcvinish et al. 2009 derive sufficient conditions on nonparametric distributions for which the consistency of the Bayes factor for testing a parametric family of distributions holds. Proving such consistency results usually comes at the cost of having a refined control on the asymptotic behavior of the marginal likelihoods that determine the Bayes Factor.

Understanding the behavior of the marginal likelihood of some data y under the Dirichlet Process mixture model, denoted by $m_{DP}(\boldsymbol{y})$, corresponds to determining asymptotic lower and upper bounds for the latter. Deriving lower bounds on marginal densities is typically done along the lines of Ghosal et al. 2000, where these bounds are used to obtain posterior concentration rates under a Dirichlet Process mixture model. There is now a large literature on posterior contraction rates in Dirichlet process mixture models, see for instance Ghosal and van der Vaart 2007, Kruijer et al. 2010, Shen et al. 2013, and Scricciolo 2014 in which a lower bound on $m_{DP}(\boldsymbol{y})$ is derived for Dirichlet Process mixtures of Gaussians.

The difficult part when assessing evidence in this setting stands in obtaining an upper bound on $m_{DP}(\boldsymbol{y})$, since it requires a refined understanding on neighborhoods of f_0, the true density from which data arise. Obtaining such an upper bound is

of interest even outside the context of testing, since it is a way to understand the behavior of credible regions in infinite dimensional models (see, e.g., Rousseau and Szabo 2020), together with proving a lower bound on posterior contraction rates as in Castillo 2008.

Obviously, practical applications of Bayes Factor consistency results for the DPM are dependent on the existence of reliable estimators of the model evidence associated to the DPM and, subsequently, to the Bayes Factor. Unfortunately, this issue still lacks a satisfactory and popular resolution. In fact, to our knowledge, only Basu and Chib 2003 directly address the issue of evidence estimation for the DPM by adapting the method of Chib 1995. SMC tools have been proposed by MacEachern et al. 1999 for beta-binomial Dirichlet Process mixtures. This idea was later generalized by Griffin 2017 and applied to a very particular kind of DPM by Tokdar and Martin 2021. However, their proposed SMC framework relies on the strong assumption that the concentration parameter M of the DP is known and fixed. Quintana and Newton 2000 give an *ad hoc* procedure in order to identify the maximum likelihood estimator of M. Neither Chib's algorithm nor SMC appear to be widely used by practitioners. Hence, it is a common practice to use the DPM without considering alternative parametric or nonparametric models, although it may not always be appropriate.

In Section 3.3, we derive an upper bound on $m_{DP}(\boldsymbol{y})$ when $f_0 \in \cup_{K\in\mathbb{N}^*}\mathfrak{M}_K$, where $\mathfrak{M}_K$ denotes the mixture models with K components $f_\theta(y)$ (Theorem 3.1). We subsequently establish the consistency of the Bayes factor comparing the parametric family of 'strongly identifiable' finite mixtures against the nonparametric location Dirichlet Process Mixture model (Corollary 3.2). This is achieved by controlling the *a priori* mass of decreasing neighborhoods of the true density f_0 (Lemma 3.3).

We propose in Section 3.4 an algorithm based on Reverse Logistic Regression (Geyer 1994) that does not require the concentration parameter M to be fixed but rather to follow a Gamma prior distribution. We show empirically that this method scales better with the amount of data than the algorithm of Basu and Chib 2003 does. We also provide a review and assessment of the ways to estimate the marginal likelihood of a nonparametric DPM model, including scenarios where n is large, which, to our knowledge, has not been done before. Moreover, we assess empirically the behavior of the Bayes Factor comparing a family of finite mixtures against a nonparametric DPM alternative.

3.2 The Dirichlet Process Mixture model

This introduction to the Dirichlet Process and the subsequent Dirichlet Process Mixture model is mainly based on the book by Hjort et al. 2010 and in particular on Chapter 2 (Ghosal 2010).

3.2.1 Notations

- Data : $\boldsymbol{y} = (y_1, \ldots, y_n)$, $n \geq 1$.
- Subset of data : $\boldsymbol{y}_{1:l} = (y_1, \ldots, y_l)$ for $l \leq n$
- $\Pi(\vartheta)$: generic notation of a prior distribution on some parameter ϑ.
- $\pi(\vartheta)$: Radon-Nikodym density of $\Pi(\vartheta)$ with respect to some measure ν (the Lebesgue measure or the counting measure, depending on the context).
- $\Pi(\vartheta|\boldsymbol{y})$: posterior distribution of some parameter ϑ given data $\boldsymbol{y}$.
- $\pi(\vartheta|\boldsymbol{y})$: Radon-Nikodym density of $\Pi(\vartheta|\boldsymbol{y})$ with respect to some measure ν (the Lebesgue measure or the counting measure, depending on the context).

3.2.2 The Dirichlet Process

The Dirichlet Process (DP) was first introduced by Ferguson 1973 as a stochastic process which realizations are probability distributions. More precisely, for Θ is a measurable space, G_0 a probability measure on Θ and some $M > 0$, the Dirichlet Process $DP(M, G_0)$ is such that for all finite partition $\{B_1, \ldots, B_K\}$ of Θ, if $P \sim DP(M, G_0)$, then

$$(P(B_1), \ldots, P(B_K)) \sim \mathcal{D}(MG_0(B_1), \ldots, MG_0(B_K)) \tag{3.1}$$

where $\mathcal{D}(\cdot)$ denote the K-dimensional Dirichlet distribution. As we will see later on, this stochastic process is commonly used in the field of Bayesian nonparametrics as a prior on the distribution of the component parameters of an infinite-dimensional mixture model, or Dirichlet Process Mixture model and we derive below the fundamental properties of the Dirichlet Process that help comprehend its attractiveness.

Expectation and variance of a realization of the DP. Deriving the first two moments of a realization of the Dirichlet Process helps understanding the specific roles of the probability measure G_0 and the so called *concentration* parameter $M > 0$ which fully characterize the DP. They can easily be computed using the trivial partition $\{B, B^c\}$ of Θ, for a measurable set B,

$$P(B) \sim \mathcal{D}(MG_0(B), M(1 - G_0(B))) \overset{d}{=} Beta(MG_0(B), M(1 - G_0(B)).$$

Therefore,

$$\mathbb{E}[P(B)] = \frac{G_0(B)}{G_0(B) + 1 - G_0(B)} = G_0(B).$$

This implies that if $\theta|P \sim P$ then

$$\begin{aligned}\mathbb{P}(\theta \in B) = \mathbb{E}(\mathbb{1}(\theta \in B)) &= \mathbb{E}[\mathbb{E}(\mathbb{1}(\theta \in B)|P)] \\ &= \mathbb{E}[\mathbb{P}(\theta \in B|P)] \\ &= \mathbb{E}[P(B)] = G_0(B)\end{aligned}$$

Hence, the marginal distribution of θ is G_0. For this reason, the latter is called *base measure*.

On the other hand,

$$\mathbb{V}[P(B)] = \frac{G_0(B)(1 - G_0(B))}{M + 1},$$

which implies that P is more concentrated around its mean as M grows. Hence the name *concentration parameter* for M. Moreover, as $M \to \infty$, $P(B) \to G_0(B)$ for any measurable set B, which implies that $P \to G$ weakly.

Conjugacy. Another interesting property of the Dirichlet Process is its conjugacy. Indeed, if one considers a sample $\{\theta_1, \ldots, \theta_n\}$ from P, then for all partition $\{B_1, \ldots, B_K\}$, equation (3.1) yields

$$\begin{aligned}\mathbb{P}(P(B_1), \ldots, P(B_K)|\theta_1, \ldots, \theta_n) &\propto \mathbb{P}(\theta_1, \ldots, \theta_n|P(B_1), \ldots, P(B_K)) \\ &\quad \times \mathbb{P}(P(B_1), \ldots, P(B_K)) \\ &= P(B_1)^{N_{B_1}+MG_0(B_1)} \times \cdots \times P(B_K)^{N_{B_K}+MG_0(B_K)}\end{aligned}$$

where $N_{B_k} = |\{i : \theta_i \in B_k\}|$.

Hence

$$P(B_1), \ldots, P(B_K)|\theta_1, \ldots, \theta_n \sim \mathcal{D}(N_{B_1} + MG_0(B_1), \ldots, N_{B_K} + MG_0(B_K)).$$

Since this is true for all partitions of Θ, then the posterior distribution of P given a sample $(\theta_1, \ldots, \theta_n)$ is again a Dirichlet Process with base measure $\frac{MG_0+\sum_{i=1}^n \delta_{\theta_i}}{M+n}$ and concentration parameter $M + n$. This writes

$$P|\theta_1, \ldots, \theta_n \sim DP\left(M + n, \frac{MG_0 + \sum_{i=1}^n \delta_{\theta_i}}{M + n}\right)$$

Predictive distribution and discreteness of realizations of the DP. A specificity of the Dirichlet Process is that its realizations are almost surely discrete distributions. To see this, consider the predictive distribution of θ_{n+1} given a sample $(\theta_1, \ldots, \theta_n) \sim P$.

$$\begin{aligned}\mathbb{P}(\theta_{n+1} \in B|\theta_1, \ldots, \theta_n) &= \int \mathbb{P}(\theta_{n+1} \in B|P)\mathbb{P}(dP|\theta_1, \ldots, \theta_n) \\ &= \int P(B)\mathbb{P}(dP|\theta_1, \ldots, \theta_n) \\ &= \mathbb{E}[P(B)|\theta_1, \ldots, \theta_n] \\ &= \frac{MG_0(B) + \sum_{i=1}^{n} \delta_{\theta_i}(B)}{M+n}.\end{aligned}$$

Hence,

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{\theta_i} \tag{3.2}$$

Thus, irrespective of the nature of the base measure G_0, sample values from P might take on the same value with non-zero probability. This implies that a realization P of the Dirichlet Process is almost surely a discrete distribution. Another interesting property induced by (3.2) is the exchangeability of the sequence $(\theta_1, \ldots, \theta_n)$. Thus it is possible to derive the full conditional distribution

$$\theta_i|\boldsymbol{\theta}_{-i} \sim \frac{M}{M+n-1}G_0 + \frac{1}{M+n-1}\sum_{i=1}^{n}\delta_{\theta_i} \tag{3.3}$$

where $\boldsymbol{\theta}_{-i} := (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$.
The next paragraph describes a useful analogy to characterize the Dirichlet process.

The Chinese Restaurant Process analogy. A useful representation of the DP can be built upon the discreteness of its realizations and the predictive distribution (3.2). Indeed, since P is almost-surely discrete, a sample $(\theta_1, \ldots, \theta_n)$ may have ties. Let us denote by $(\theta_1^*, \ldots, \theta_{K_+}^*)$ the $K_+ \leq n$ unique values in the sample. Then a new value θ_{n+1} is sampled according to the distribution

$$\theta_{n+1}|\theta_1^*, \ldots, \theta_K^* \sim \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{k=1}^{K}N_k\delta_{\theta_k^*} \tag{3.4}$$

with $N_k = \sum_{i=1}^{n}\mathbb{1}(\theta_i = \theta_k^*)$, where we have simply used the predictive distribution (3.2).
This is conveniently formalized by the so-called Chinese Restaurant Process (CRP, Aldous 1985), which is a discrete-time stochastic process, best described with the analogy of sitting customers arriving one by one in a restaurant with an infinite number of tables. That is, in this restaurant analogy, a new customer θ_{n+1} sits at an existing table with probability $N_k/(M+n)$, while it sits at a new table with probability $M/(M+n)$. Once the customer is seated, distribution (3.4) is updated according to the table allocation of θ_{n+1} before proceeding with seating the next new customer, θ_{n+2}.

This representation helps better understand the 'rich gets richer' phenomenon at play with the Dirichlet Process. A new customer will seat at an existing table

with probability proportional to the number of customers already seated at this table. In fact, the expected number of tables as n grows to infinity is $O(M \log n)$. Indeed, since for all $i = 1, \ldots, n$ the probability that customer i sits at a new table is $M/(M+i-1)$. Hence, the expected number of table is given by $\sum_{i=1}^{n} M/(M+i-1)$ which is upper-bounded by MH_n where H_n is the n-th harmonic number.

The CRP is also useful for understanding the partitioning structure induced by the Dirichlet Process. The unique values $(\theta_1^*, \ldots, \theta_K^*)$ among $(\theta_1, \ldots, \theta_n)$ define indeed a partition on the space $[n]$, which makes the DP particularly attractive as a prior on the distribution of the parameters of the components of a mixture model, as we shall see in the next section.

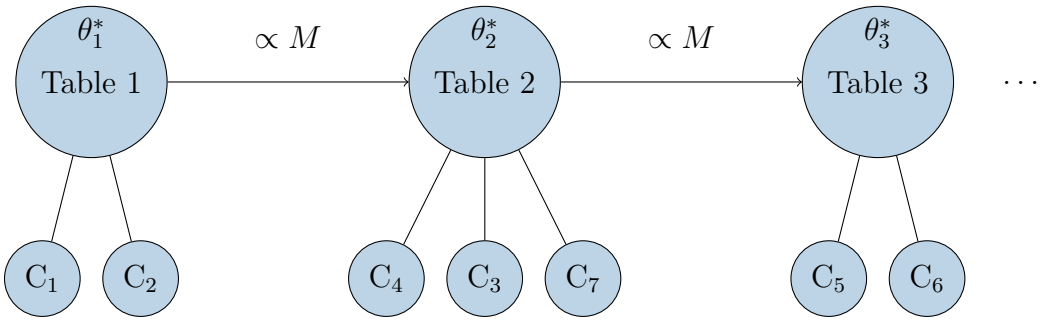


Figure 3.1: Illustration of the Chinese Restaurant Process. C_i stands for customer i.

Stick-breaking representation of realizations of the DP. Sethuraman 1994 gives the following convenient representation of realizations of a DP as discrete distributions on Θ with infinitely many atoms. It is derived as

$$\begin{aligned} \theta_1, \theta_2, \ldots &\overset{i.i.d}{\sim} G_0 \\ V_1, V_2, \ldots &\overset{i.i.d}{\sim} Beta(1, M) \\ \varpi_k &= V_k \prod_{j=1}^{k-1} (1 - V_j), \quad \text{for } k = 1, 2, \ldots \\ P &= \sum_{k=1}^{\infty} \varpi_k \delta_{\theta_k} \end{aligned} \tag{3.5}$$

The induced distribution on the weights $\boldsymbol{\varpi}$ is usually denoted $GEM(M)$ and is called the stick-breaking distribution. Indeed, if one assimilates the segment $[0, 1]$ with a stick of length 1 that is first broken at $\varpi_1 = V_1$, a proportion V_2 of the remaining length of the stick $(1 - V_1)$ is then broken, and so on, as illustrated on Figure 3.2. This mechanism ensures that $\sum_{k=1}^{\infty} \varpi_k = 1$ a.s.

The representation of P as a discrete distribution given in (3.5) makes the Dirichlet Process an intuitively convenient choice of prior for the parameters of an 'infinite' mixture model that we shall introduce in the next section.

Figure 3.2: Illustration of the stick breaking process for $V_1 = 0.2, V_2 = 0.4, V_3 = 0.3$ and the resulting weights ϖ_1, ϖ_2 and ϖ_3 represented on a 'stick' of length one.

3.2.3 The Dirichlet Process Mixture model - specification and posterior inference

First formalized by Ferguson 1983, the Dirichlet Process Mixture model (DPM) can be specified as

$$\begin{aligned} y_i|\theta_i &\overset{i.i.d}{\sim} F(y_i|\theta_i) \quad \text{for } i = 1, \dots, n \\ \theta_i|P &\sim P \\ P|G_0, M &\sim DP(M, G_0) \\ M &\sim \Pi_M \end{aligned} \tag{3.6}$$

for a collection of data points $\boldsymbol{y} = (y_1, \dots, y_n)$ where $F(\cdot|\theta)$ is a distribution supported on $\mathbb{R}^d$ with density with respect to the Lebesgue measure $f(\cdot|\theta)$. The discrete nature of realizations of the Dirichlet Process P enables ties in the parameters and therefore a clustering of the data $\boldsymbol{y}$.

An alternative specification to (3.6) that truly highlights the mixture nature of the DPM makes use of the stick-breaking representation (3.5) and reads

$$\begin{aligned} y_i|z_i, \boldsymbol{\theta} &\overset{i.i.d}{\sim} F(y_i|\theta_{z_i}) \quad \text{for } i = 1, \dots, n \\ \theta_1, \theta_2, \dots &\overset{i.i.d}{\sim} G_0 \\ \mathbb{P}(z_i = k|\boldsymbol{\varpi}) &= \varpi_k \\ \varpi_1, \varpi_2, \dots|M &\sim GEM(M) \\ M &\sim \Pi_M \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_i)_{i=1}^{\infty}$.
Equivalently, since $P = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ a.s, the likelihood of data $\boldsymbol{y}$ given P is easily derived as

$$f_P(\boldsymbol{y}) := p(\boldsymbol{y}|P) = \prod_{i=1}^{n} \int_{\Theta} f(y_i|\theta) dP(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{\infty} \varpi_k f(y_i|\theta_k) \tag{3.7}$$

The link between the likelihood (3.7) and that of a finite mixture is immediate and explains the commonly used 'infinite mixture' term for denoting the DPM. In fact, the DPM arises as the limit of a $K-$dimensional finite mixture with Dirichlet prior $\mathcal{D}(M/K, \dots, M/K)$, as $K \to 0$.

The likelihood (3.7) is not convenient to work with due to its intractability. It is common to consider instead the likelihood augmented with the latent allocation

variables $\boldsymbol{z} = (z_1, \ldots, z_n)$ given by

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{z}) &= \int p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}) \\ &= \int \prod_{i=1}^{n} p(y_i|z_i, \boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}) \\ &= \prod_{k=1}^{K_+} \int \prod_{i:z_i=k} f(y_i|\theta_k^*) G_0(d\theta_k^*) \\ &= \prod_{k=1}^{K_+} m_k(\boldsymbol{z}) \end{aligned} \tag{3.8}$$

where $m_k(\boldsymbol{z}) := \int \prod_{i:z_i=k} f(y_i|\theta_k^*) G_0(d\theta_k)$ is the marginal likelihood of data allocated to component k and K_+ is the number of non-empty clusters induced by the allocation vector $\boldsymbol{z}$. This quantity is available in closed form for a conjugate prior G_0 to the likelihood function f.

From the predictive distribution (3.2) one can derive the prior induced on the allocation vector $\boldsymbol{z}$ that writes

$$\pi(\boldsymbol{z}|M) = \pi(z_1, \ldots, z_n|M) = \frac{\Gamma(M)}{\Gamma(M+n)} M^{K_+} \prod_{k=1}^{K_+} \Gamma(N_k). \tag{3.9}$$

Notice that equation (3.9) implies that the sequence of random variables $\boldsymbol{z} = (z_1, \ldots, z_n)$ is exchangeable under the DP prior.

The posterior distribution on the allocation vector can then be derived up to a constant as

$$\pi(\boldsymbol{z}|\boldsymbol{y}, M) \propto p(\boldsymbol{y}|\boldsymbol{z}) \pi(\boldsymbol{z}|M)$$

On the other hand, full conditional posteriors on $\boldsymbol{\theta}$ can be obtained up to a proportionality constant by using equation (3.3) and multiplying by $f(y_i|\theta_i)$

$$\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{y}, M \propto M f(y_i|\theta_i) G_0(\theta_i) + \sum_{k=1}^{K_+^{-i}} N_k^{-i} f(y_i|\theta_k^*) \delta_{\theta_k^{*-i}}$$

where the quantities with a superscript $-i$ denote the appropriate quantities once observation y_i is ignored.
Notice that this can be rewritten as

$$\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{y}, M \propto M \int_\Theta f(y_i|\theta_i) G_0(d\theta_i) \times \Pi(\theta_i|y_i) + \sum_{k=1}^{K_+^{-i}} N_k^{-i} f(y_i|\theta_k^{*-i}) \delta_{\theta_k^{*-i}}$$

which defines a Gibbs Sampling strategy where θ_i is either given the value of one of the unique $\boldsymbol{\theta}^{*-i}$ with probability proportional to $N_k^{-i} f(y_i|\theta_k^*)$, or assigned a new value of theta from the posterior distribution given y_i only with probability proportional to $M \int_\Theta f(y_i|\theta_i) G_0(d\theta_i)$. Another interpretation is to first sample the allocation z_i

according to its full conditional

$$\Pi(z_i|\boldsymbol{z}_{-i}, \boldsymbol{\theta}^{*-i}, \boldsymbol{y}, M) \propto \begin{cases} M \int_\Theta f(y_i|\theta) G_0(d\theta) & \text{if } z_i = K_+^{-i} + 1 \\ N_{z_i}^{-i} f(y_i|\theta_{z_i}^{*-i}) & \text{if } z_i = 1, \ldots, K_+^{-i} \end{cases} \tag{3.10}$$

before sampling θ_i as

$$\Pi(\theta_i|z_i, \boldsymbol{z}_{-i}, \boldsymbol{\theta}^{*-i}, \boldsymbol{y}) = \begin{cases} \Pi(\theta_i|y_i) & \text{if } z_i = K_+^{-i} + 1 \\ \delta_{\theta_{z_i}^{*-i}} & \text{if } z_i = 1, \ldots, K_+^{-i} \end{cases} \tag{3.11}$$

This is summarized in Algorithm 9 below.

Algorithm 9 : Gibbs Sampler for conditionally conjugate DPM

1 At step t,
2 **for** $i = 1, \ldots, n$ **do**
3 Sample $z_i^{(t)}|\boldsymbol{z}_{-i}^{(t)}, (\boldsymbol{\theta}^{*-i})^{(t)}, \boldsymbol{y}, M$ according to (3.10);
4 Sample $\theta_i|z_i, \boldsymbol{z}_{-i}, \boldsymbol{\theta}^{*-i}, \boldsymbol{y}$ according to (3.11);
5 **end**
6 Return $(\{\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}\}_t)$ distributed according to $\Pi(\boldsymbol{\theta}, \boldsymbol{z}|y, M)$

The parameters $\boldsymbol{\theta}$ can hinder good mixing within the above Gibbs sampler. Thus, if clustering is the only purpose of inference, it is common to integrate with respect to $\Pi(\boldsymbol{\theta}^{*-i}|\boldsymbol{y})$ which yields

$$\Pi(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{y}, M) \propto \begin{cases} M \int_\Theta f(y_i|\theta) G_0(d\theta) & \text{if } k = K_+^{-i} + 1 \\ N_k^{-i} \int_\Theta f(y_i|\theta_k^{*-i}) d\Pi(\theta_k^{*-i}|\boldsymbol{y}_k^{-i}) & \text{if } k = 1, \ldots, K_+^{-i} \end{cases} \tag{3.12}$$

where $\boldsymbol{y}_k^{-i} = \{y_l : l \neq i \text{ and } z_l = k\}$ and

$$\Pi(\theta|\boldsymbol{y}_k^{-i}) \propto G_0(\theta) \prod_{l \neq i, s_l = k} F(y_l|\theta)$$

is the posterior distribution computed on data allocated to group k, excluding y_i, which is analytical for a conjugate pair G_0 and F.
Hence equation (3.12) can be rewritten as

$$\Pi(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{y}, M) \propto \begin{cases} M \int_\Theta f(y_i|\theta) G_0(d\theta) & \text{if } k = K_+^{-i} + 1 \\ N_k^{-i} m_k(\boldsymbol{z}^{-i} \cup \{z_i = k\}) / m_k(\boldsymbol{z}^{-i}) & \text{if } k = 1, \ldots, K_+^{-i} \end{cases} \tag{3.13}$$

where $m_k(\boldsymbol{z})$ is defined as in equation (3.8).

This defines the so called *collapsed* Gibbs Sampler (Neal 2000) as detailed below.

Algorithm 10 : Collapsed Gibbs Sampler for conditionally conjugate DPM

1 At step t,
2 **for** $i = 1, \dots, n$ **do**
3 | Sample $z_i^{(t)} | \boldsymbol{z}_{-i}^{(t)}, \boldsymbol{y}, M$ according to (3.13);
4 **end**
5 Return $\{\boldsymbol{z}^{(t)}\}_t$ distributed according to $\Pi(\boldsymbol{z}|y, M)$

As indicated in our model specification (3.6), it is common practice to assume a prior distribution Π_M on M. In fact, when estimating the number of clusters present in a dataset, Ascolani et al. 2023 shows that a well-chosen prior on M can lead to a consistent estimation procedure while Miller and Harrison 2013 demonstrate that it is not true when M is held fixed.

However, computing the full conditional posterior distribution on M to perform Gibbs sampling is not trivial. Escobar and West 1995 give a convenient data-augmentation strategy to be able to infer M within the Gibbs sampling framework. Using the posterior on K_+ given in Antoniak 1974,

$$\begin{aligned}\pi(K_+|M, \boldsymbol{y}) &= c_n(K_+) n! M^{K_+} \frac{\Gamma(M)}{\Gamma(M+n)} \\ &= c_n(K_+) n! M^{K_+ - 1} \frac{\beta(M+1, n)(M+n)}{\Gamma(n)}\end{aligned}$$

where $c_n(K_+)$ does not depend on M and $\beta(\cdot, \cdot)$ denotes the usual Beta function, the posterior full conditional on M can be written

$$\begin{aligned}\pi(M|K_+, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}) &= \pi(M|K_+, \boldsymbol{y}) \\ &\propto \pi(M)\pi(K_+|M, \boldsymbol{y}) \\ &\propto \pi(M) M^{K_+ - 1}(M+n) \int_0^1 x^M (1-x)^{n-1} dx.\end{aligned}$$

Thus, the posterior full conditional of M can be seen as the marginal distribution of the joint distribution on (M, η) for some random variable η such that

$$\pi(M, \eta|K_+, \boldsymbol{y}) \propto \pi(M) M^{K_+ - 1}(M+n)\eta^M (1-\eta)^{n-1}$$

for $0 < \eta < 1$.
Now note that for a Gamma prior $\Gamma(a, b)$ on M, given η,

$$\begin{aligned}\pi(M|\eta, K_+, \boldsymbol{y}) &\propto M^{a+K_+-2} \exp\{-(b - \log\eta)M\}(M+n) \\ &\propto M^{a+K_+-1} \exp\{-(b-\log\eta)M\} + nM^{a+K_+-2}\exp\{-(b-\log\eta)M\} \\ &\propto \frac{(b-\log\eta)^{a+K_+}}{\Gamma(a+K_+)} M^{a+K_+-1}\exp\{-(b-\log\eta)M\} \\ &\quad + \frac{n(b-\log\eta)}{a+K_+-1}\frac{(b-\log\eta)^{a+K_+-1}}{\Gamma(a+K_+-1)} M^{a+K_+-2}\exp\{-(b-\log\eta)M\}\end{aligned}$$

which is a mixture of two Gamma distributions,

$$\Pi(M|\eta, K_+, \boldsymbol{y}) = \omega\Gamma(a + K_+, b - \log\eta) + (1 - \omega)\Gamma(a + K_+ - 1, b - \log\eta) \quad (3.14)$$

with $\omega = (a + K_+ - 1)/\{a + K_+ - 1 + n(b - \log\eta)\}$.
Trivially,

$$\Pi(\eta|M, \boldsymbol{y}) = \mathcal{B}eta(M + 1, n).$$

Hence the following algorithm gives a full Gibbs sampling strategy when a $\Gamma(a, b)$ prior is set on the concentration parameter.

Algorithm 11 : Collapsed Gibbs Sampler for conditionally conjugate DPM and $M \sim \Gamma(a, b)$ a priori

1 At step t,
2 **for** $i = 1, \ldots, n$ **do**
3 Sample $z_i^{(t)}|\boldsymbol{z}_{1:i-1}^{(t)}, \boldsymbol{z}_{i+1:n}^{(t-1)}, \boldsymbol{y}, M^{(t)}$ according to (3.13);
4 **end**
5 Sample $\eta^{(t)}|K_+^{(t)}, M^{(t-1)}, \boldsymbol{y} \sim \mathcal{B}eta(1 + M^{(t-1)}, n)$;
6 Sample $M^{(t)}|K_+^{(t)}, \eta^{(t)}, \boldsymbol{y}$ using the mixture (3.14);
7 Return $\{\boldsymbol{z}^{(t)}, M^{(t)}\}_t$ distributed according to $\Pi(\boldsymbol{z}, M|y)$

While these Gibbs sampling approaches are well-established, there still remains computational challenges when working with the DPM. For instance the issue of Bayesian model selection and assessment, which in practice heavily depends on the availability of model evidence estimation techniques. However, this issue remains largely unexplored so far in Bayesian non-parametrics. To the best of our knowledge, only Basu and Chib 2003 directly addresses the problem of estimating the marginal likelihood of a DPM. For this reason, this algorithm has never been properly assessed and compared to alternative methods.

From a theoretical perspective, deriving an upper-bound on the marginal likelihood in a Dirichlet Process mixture model remains an open challenge. We provide such a result in the next Section. Our main contribution is the derivation of an asymptotic upper bound on $m_{DP}(\boldsymbol{y})$ (Theorem 3.1). This is done by controlling the a priori mass of decreasing L_1-neighborhoods of the true mixture density f_0 (Lemma 3.3). A consequence is the consistence of the Bayes Factor comparing a parametric model against a DPM alternative (Corollary 3.2).

In Section 3.4, we propose an empirical illustration of the above-mentioned theoretical results. Doing so requires reliable estimates to the model evidence $m_{DP}(\boldsymbol{y})$. We derive a new approach based on Geyer's Reverse Logistic Regression (RLR) trick (Geyer 1994) and compare its performance with Chib's algorithm for the DPM.

3.3 Asymptotics of the evidence associated to the DPM

In this section we study the asymptotic behaviour of the marginal likelihood $m_{DP}(\boldsymbol{y}) = \int f_P(\boldsymbol{y}) d\Pi(P)$ when Π is the Dirichlet Process $DP(M, G_0)$ for some data $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $y_i \overset{iid}{\sim} f_0$, a probability density on $\mathbb{R}^d$. Understanding the behaviour of $m_{DP}(\boldsymbol{y})$ corresponds to determining asymptotic lower and upper bounds for $m_{DP}(\boldsymbol{y})$. Deriving lower bounds on marginal densities is typically done using the technique of Ghosal et al. 2000, Lemma 8.1 for instance, where it is is used to derive posterior concentration rates under a Dirichlet Process mixture model. There is now a large literature on posterior contraction rates in Dirichlet process mixture models, see for instance Ghosal and van der Vaart 2007, Kruijer et al. 2010, Shen et al. 2013, and Scricciolo 2014 in which a lower bound on $m_{DP}(\boldsymbol{y})$ is derived.

The difficult part when assessing evidence in this setting stands in obtaining an upper bound on $m_{DP}(\boldsymbol{y})$, since it requires a refined understanding on neighbourhoods of f_0. In the following section we concentrate on deriving such an upper bound when $f_0 \in \cup_{K \in \mathbb{N}^*} \mathfrak{M}_K$, where $\mathfrak{M}_K$ denotes the model of mixtures with K densities $f_\theta(y)$. Indeed an important application of such an upper bound is in the goodness of fit test (or test for the number of components) $H_0 : f_0 \in \mathfrak{M}_K$, versus $H_1 : f_0 \notin \mathfrak{M}_K$, to prove that the Bayes factor is consistent under the null, i.e that $m_K(\boldsymbol{y})/m_{DP}(\boldsymbol{y}) \to \infty$ in probability under f_0.

3.3.1 Main result

In this section we assume that $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$ where $\varpi_j^0 > 0$ for all j and $\theta_j^0 \neq \theta_i^0$ for all $i \neq j$, resulting in

$$f_0 = f_{P_0} := \int_\Theta f(y|\theta) P_0(d\theta) = \sum_{j=1}^{K_0} \varpi_j^0 f(y|\theta_j^0).$$

We also consider the following regularity assumptions on the distribution $f(y|\theta)$.

Assumption A1 [Regularity] For all $y \in \mathcal{Y} \subset \mathbb{R}^d$, the function $\theta \to f(y|\theta)$ is twice continuously differentiable and there exist $H_1 \in L^2(\mathbb{R}^d)$, $H_2 \in L^1(\mathbb{R}^d)$, $H_3 \in L^1(\mathbb{R}^d)$, and $\delta_0, \delta > 0$ such that for all $j = 1, \cdots, K_0$,

$$\begin{aligned}
\sup_{\theta \in \Theta} \|\nabla f(y|\theta)\| &\leq H_1(y), \\
\sup_{\|\theta - \theta_j^0\| \leq \delta_0} \|D^2 f(y|\theta)\| &\leq H_2(y) \\
\|D^2 f(y|\theta) - D^2 f(y|\theta_j^0)\| &\leq H_3(y) \|\theta - \theta_j^0\|^\delta, \ \forall \|\theta - \theta_0\| \leq \delta_0,
\end{aligned}$$

where ∇f and $D^2 f$ denote respectively the gradient and the Hessian matrix of f.

We then consider a strong identifiability assumption, similar to the one used in Chen 1995 or Nguyen 2013. Denote by S_d^+ the set of symmetrical semi-definite matrices of dimension d.

Assumption A2 [Strong identifiability] For all $\epsilon > 0$ and all ν_0 non null measures on $A_0 = [\cup_{j=1}^{K_0} B(\theta_j^0, \epsilon)]^c$ satisfying $\nu_0(A_0) \leq 1$ and all $\alpha_0 \geq 0$, $\alpha_j \in \mathbb{R}$, $\beta_j \in \mathbb{R}^d$ and $\gamma_j \in S_d^+$, $j = 1, \cdots, K_0$,

$$\alpha_0 f_{\nu_0}(y) + \sum_{j=1}^{K_0} [\alpha_j f_{\theta_j^0} + \beta_j \nabla f_{\theta_j^0} + tr(D^2 f_{\theta_j^0} \gamma_j)] = 0 \quad \Leftrightarrow \alpha_0 = \alpha_j = 0,\ \beta_j = 0,\ \gamma_j = 0$$

where $f_\lambda = \int f_\theta(x)\lambda(d\theta)$ for all finite measure λ on $\mathbb{R}^d$.
Assumption A3 For all x, $f_\theta(x)$ converges to 0 at $\bar{\Theta} \cap \Theta^c$, where $\bar{\Theta}$ is the closure of Θ and $\sup_\theta \|f_\theta(\cdot)\|_\infty < \infty$.
Assumption A4 The Dirichlet Process base measure G_0 has positive density on $\mathbb{R}^d$ and verifies: $\int_\Theta e^{a_1\|\boldsymbol{\theta}\|^{a_2}} dG_0(\boldsymbol{\theta}) < \infty$, for some $a_1, a_2 > 0$
Assumption A5 The prior Π_M on M has support on $[\zeta, \infty)$ for some $\zeta > 0$ and is such that

$$\int_\zeta^\infty \exp\{M(\delta_0 + (1-\delta_1)\log K_0)\}\, d\Pi_M(M) < \infty$$

where $\delta_0, \delta_1 > 0$ can be chosen arbitrarily small.

Remark 3.1. *The cornerstone of the proof of Lemma 3.3 is the existence and boundedness of the density function of the Dirichlet Process random mean defined as $\mu(P) := \int \theta dP(\theta)$ for P a realization of the Dirichlet Process $DP(M, G_0)$. Feigin and Tweedie 1989 show that if $\int_\Theta \log(1 + \|\boldsymbol{\theta}\|) dG_0(\boldsymbol{\theta}) < \infty$ then the random mean $\mu(P)$ exists. Thus, Assumption **A4** clearly ensures the existence of $\mu(P)$ and is also necessary to the proof of Theorem 3.1.*

Remark 3.2. *Assumption **A5** is a sufficient but not a necessary condition for the proof of Lemma 3.3. The truncation on the support of the concentration parameter M can be understood as a way for the Bayes factor to efficiently discriminate between the DPM and finite mixture models. Indeed, as M goes to 0, the number of non-empty clusters given by the DPM is close to 1, making the distinction between the two models difficult. One prior distribution satisfying this assumption is the truncated Gamma distribution with rate $\beta > \delta_0 + (1-\delta_1)\log K_0$.*

The next Theorem shows that under Assumptions **A1**–**A5** and if f_{P_0} is a mixture with K_0 components, then $m_{DP}(\boldsymbol{y})$ is bounded from above by $n^{-(K_0 - 1 + dK_0 + t)/2}$ for some $0 < t < \zeta$.

Theorem 3.1. *Assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ are i.i.d observations from $f_0 := f_{P_0}$ with $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$ and that Assumptions **A1**–**A3** are satisfied. Denote by Π the joint prior distribution on (P, M) where $P|M \sim DP(M, G_0)$ and $M \sim \Pi_M$, such that Π verifies Assumptions **A4**–**A5**. Then for any $0 < t < \zeta$,*

$$P_{f_0}\left(m_{DP}(\boldsymbol{y}) > n^{-(K_0 - 1 + dK_0 + t)/2}\right) \overset{n\to\infty}{\longrightarrow} 0$$

where

$$m_{DP}(\boldsymbol{y}) := \int f_P(\boldsymbol{y})/f_0(\boldsymbol{y}) d\Pi(P)$$

A consequence of Theorem 3.1 is the convergence of the Bayes factor:

Corollary 3.2. *If the DP prior verifies **A4**–**A5** and the prior on $\mathfrak{M}_{K_0}$ is defined as in (2.5), then*

*(i) If $f_{P_0} \in \mathfrak{M}_{K_0}$ satisfies Assumptions **A1**–**A3**, then $m_{K_0}(\mathbf{y})/m_{DP}(\mathbf{y}) \to \infty$ under f_{P_0}.*

(ii) Moreover for all $K > K_0$, if the prior on $\mathfrak{M}_K$ is defined as in (2.5) with Dirichlet hyperparameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$ for some $\alpha > 0$, if either $\alpha < d/2 \wedge \zeta/(K - K_0)$, or $d/2 < \alpha \wedge \zeta/(K - K_0)$, then $m_K(\mathbf{y})/m_{DP}(\mathbf{y}) \to \infty$ under f_{P_0}.

(iii) If $\inf_{f_P \in \mathfrak{M}_K} KL(f_{P_0}, f_P) > 0$ and the DP prior verifies $\Pi_{DP}(KL(f_{P_0}, f_P) \leq \epsilon) > 0$ for all $\epsilon > 0$, then $m_K(\mathbf{y})/m_{DP}(\mathbf{y}) \to 0$ under f_{P_0}.

Remark 3.3. *The sufficient conditions on the truncation parameter ζ of the prior on M in Corollary 3.2 for the overfitted case can be understood as a way to force the concentration parameter of the Dirichlet Process not to be too small a priori. This indeed favors larger numbers of clusters induced by the DPM and in turn makes the finite mixture alternative with $K > K_0$ a better fit. Notice that ζ must grow as the number of extra components $K - K_0$ increases.*

Remark 3.4. *By exploiting the proof of Proposition 3.10 of Gassiat and Van Handel 2014 (see Proposition 3.12) Assumption **A2** is established in location mixtures, i.e. for models in the form $f_\theta(y) = f(y - \theta)$ and Theorem 3.1 is valid under Assumptions **A1, A3, A4, A5**.*

Remark 3.5. *A sufficient condition for Assumption **A4** to be true is that distribution G_0 has a finite first moment, which includes Gaussian distributions for instance. Note that although the Cauchy distribution has no finite expectation, it does verify Assumption **A4**.*

The following Lemma is at the core of the proof of Theorem 3.1 and provides an upper bound on decreasing L_1 neighborhoods of the true density f_0.

Lemma 3.3. *Assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ are i.i.d observations from $f_0 := f_{P_0}$ with $P_0 = \sum_{j=1}^{K_0} \varpi_j^0 \delta_{\theta_j^0}$ and that Assumptions **A1**–**A3** are satisfied. Denote by Π the joint prior distribution on (P, M) where $P|M \sim DP(M, G_0)$ and $M \sim \Pi_M$, such that Π verifies Assumptions **A4**–**A5**. Then for any sequence δ_n such that $\delta_n \overset{n\to\infty}{\longrightarrow} 0$,*

$$\Pi\left(\|f_P - f_0\|_1 < \delta_n\right) \lesssim \delta_n^{K_0 - 1 + dK_0 + \zeta - \varepsilon}$$

where $\varepsilon > 0$ can be chosen as close to 0 as desired.

Proof. We wish to prove that

$$\Pi(\|f_P - f_0\|_1 \leq \delta_n) = o(n^{-D(K_0)/2 - t/2}).$$

for some $\zeta > t > 0$, where ζ is the truncation on the support of M given by Assumption **A5**. Using a similar idea to Lemma 3.8 of Gassiat and Van Handel 2014,

we construct balls $A_j := B(\theta_j^0, \epsilon)$ for all $j = 1, \ldots, K_0$ around the true parameters θ_j^0 with radius

$$\epsilon \leq \min\{\|\theta_i^0 - \theta_j^0\|/4; i \neq j \leq K_0\},$$

so that the balls A_j are disjoint. We then define

$$p_j := P(B(\theta_j^0, \epsilon)) = \sum_{i:\theta_i \in A_j} \varpi_i$$

and

$$A_0 := \left(\cup_{j=1}^{K_0} A_j\right)^C$$

so that $\{A_0, \ldots, A_{K_0}\}$ is a partition of Θ.

Using this partition, we group the atoms of the discrete distribution P from the Dirichlet Process with their 'closest' corresponding true parameter θ_j^0 as following :

$$\begin{aligned}
\|f_0 - f_P\|_1 &= \left\| \sum_{i:\theta_i \in A_0} \varpi_i f_{\theta_i} + \sum_{j=1}^{K_0} \left\{ \sum_{i:\theta_i \in A_j} \varpi_i f_{\theta_i} \right\} + \sum_{j=1}^{K_0} \left\{ \sum_{i:\theta_i \in A_j} \varpi_i f_{\theta_j^0} \right\} \right. \\
&\quad \left. - \sum_{j=1}^{K_0} \left\{ \sum_{i:\theta_i \in A_j} \varpi_i f_{\theta_j^0} \right\} - \sum_{j=1}^{K_0} \varpi_j^0 f_{\theta_j^0} \right\|_1 \\
&= \left\| \sum_{i:\theta_i \in A_0} \varpi_i f_{\theta_i} + \sum_{j=1}^{K_0} \left\{ \sum_{i:\theta_i \in A_j} \varpi_i f_{\theta_j^0} - \varpi_j^0 f_{\theta_j^0} \right\} \right. \\
&\quad \left. + \sum_{j=1}^{K_0} \left\{ \sum_{i:\theta_i \in A_j} \varpi_i (f_{\theta_i} - f_{\theta_j^0}) \right\} \right\|_1 \\
&= \left\| \int_{A_0} f_\theta dP(\theta) + \sum_{j=1}^{K_0} \left\{ \left(p_j - \varpi_j^0\right) f_{\theta_j^0} + \int_{A_j} \left(f_\theta - f_{\theta_j^0}\right) dP(\theta) \right\} \right\|_1
\end{aligned}$$

A Taylor expansion of f_θ around θ_j^0 yields

$$\begin{aligned}
&= \left\| \int_{A_0} f_\theta dP(\theta) + \sum_{j=1}^{K_0} \left\{ \left(p_j - \varpi_j^0\right) f_{\theta_j^0} + \int_{A_j} \left(\theta - \theta_j^0\right)^T \nabla f_{\theta_j^0} dP(\theta) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} \int_{A_j} \left(\theta - \theta_j^0\right)^T D^2 f_{\theta_j^0} \left(\theta - \theta_j^0\right) dP(\theta) \right\} + R \right\|_1 \\
&:= \|\Delta(P) + R\|_1
\end{aligned}$$

where

$$\begin{aligned}\|R\|_1 &\leq \sum_{j=1}^{K_0} \int_{A_j} \left\| \sup_{\theta':\|\theta'-\theta_0^j\|\leq\|\theta-\theta_j^0\|} \left| \left(\theta-\theta_j^0\right)^T \left[D^2 f_{\theta'} - D^2 f_{\theta_j^0}\right] \left(\theta-\theta_j^0\right) \right| \right\|_1 dP(\theta) \\ &\leq \sum_{j=1}^{K_0} \int_{A_j} \left\| \|\theta-\theta_j^0\|^2 \sup_{\theta':\|\theta'-\theta_0^j\|\leq\|\theta-\theta_j^0\|} \left\| D^2 f_{\theta'} - D^2 f_{\theta_j^0} \right\| \right\|_1 \\ &\leq \sum_{j=1}^{K_0} \|H_3\|_1 \int_{A_j} \left\|\theta-\theta_j^0\right\|^{2+\delta} dP(\theta).\end{aligned}$$

By Lemma 3.7 in Appendix, there exists a constant $c(f_0)$ depending only on f_0 such that

$$\begin{aligned}\|\Delta(P)\|_1 \geq c(f_0) \Bigg[P(A_0) + \sum_{j=1}^{K_0} \Bigg\{ \left|p_j - \varpi_j^0\right| + \left\| \int_{A_j} \left(\theta-\theta_j^0\right) dP(\theta) \right\| \\ + \int_{A_j} \left\|\theta-\theta_j^0\right\|^2 dP(\theta) \Bigg\} \Bigg] \\ := c(f_0) N(P)\end{aligned}$$

Then if $\|f_0 - f_P\|_1 \leq \delta_n$

$$\begin{aligned}\delta_n \geq \|\Delta(P) + R\|_1 &\geq |\|\Delta(P)\|_1 - \|R\|_1| \\ &\geq \|\Delta(P)\|_1 - \|R\|_1 \\ &\geq c(f_0) N(P) - \|R\|_1\end{aligned}$$

Hence

$$N_n(P) c(f_0) \leq \delta_n + \sum_{j=1}^{K_0} \|H_3\|_1 \int_{B(\theta_j^0,\epsilon)} \|\theta-\theta_j^0\|^{2+\delta} dP(\theta)$$

and choosing ϵ small enough gives

$$N_n(P) \leq \frac{2\delta_n}{c(f_0)} \lesssim \delta_n.$$

We now bound from above $\Pi(N_n(P) \leq \delta_n | M)$. Consider

$$P_j = \frac{P \mathbb{1}_{A_j}}{p_j}$$

and

$$p_0 = P(A_0)$$

then under the Dirichlet Process prior $(p_0, \cdots, p_{K_0})$ and $P_1, \cdots, P_{K_0}$ are mutually independent and for all $j = 1, \ldots, K_0$, $P_j \sim DP(MG_j, G_{0,j})$ where $G_j := G_0(B(\theta_j^0, \epsilon))$

and $G_{0,j} := G_0(\cdot \cap B(\theta_j^0, \epsilon))$. Hence

$$\begin{aligned}\Pi(N_n(P) \leq \delta_n | M) \leq & \Pi(p_0 \leq \delta_n, |p_j - \varpi_j^0| \leq \delta_n, \ \forall j = 1, \dots K_0) \\ & \times \prod_{j=1}^{K_0} \Pi\left(\left\| \int_{B(\theta_j^0, \epsilon)} (\theta - \theta_j^0) dP_j(\theta) \right\| \leq \delta_n \right)\end{aligned}$$

Let Π_j be the marginal Dirichlet Process prior distribution of $\int \theta dP_j(\theta)$ and let π_j denote its density w.r.t the Lebesgue measure. By Lemma 3.6 in Appendix, we know that π_j is continuous and we have that

$$\begin{aligned}\Pi\left(\left\| \int_{B(\theta_j^0, \epsilon)} (\theta - \theta_j^0) dP_j(\theta) \right\| \leq \delta_n \right) &= \Pi\left(\left\| \int_{B(\theta_j^0, \epsilon)} \theta dP_j(\theta) - \theta_j^0 \right\| \leq \delta_n \right) \\ &= \Pi_j(B(\theta_j^0, \delta_n)) \\ &\leq \sup_{\|x - \theta_j^0\| \leq \delta_n} \pi_j(x) (\delta_n)^d \\ &\lesssim \left(\frac{\Gamma(M)}{\Gamma(1+M)} \right)^{d+1} \delta_n^d\end{aligned}$$

where we have used the constant in M found in the proof of Lemma 3.5 in Appendix for the last inequality. Hence,

$$\begin{aligned}\Pi(N_n(P) \leq \delta_n | M) \lesssim & \frac{\Gamma(M)}{\prod_{j=0}^{K_0} \Gamma(MG_j)} \int_{S_n} \prod_{j=1}^{K_0} p_j^{MG_j - 1} (1 - \sum_{j=1}^{K_0} p_j)^{MG_0 - 1} dp_1, \dots, dp_{dK_0} \\ & \times M^{-K_0(d+1)} \delta_n^{dK_0}\end{aligned}$$

where $S_n := \Delta_{K_0} \cap \{(p_1, \dots, p_{K_0} : |p_j - \varpi_j^0| \leq \delta_n, \quad \forall j = 1, \dots K_0\}$, with Δ_{K_0} the K_0-dimensional simplex.

$$\begin{aligned}& \leq \frac{\Gamma(M)}{\prod_{j=0}^{K_0} \Gamma(MG_j)} \prod_{j=1}^{K_0} (\varpi_j^0 - \delta_n)^{-1} \int_{S_n} (1 - \sum_{j=1}^{K_0} p_j)^{MG_0 - 1} dp_1 \dots dp_{K_0} \\ & \times M^{-K_0(d+1)} \delta_n^{K_0 d}\end{aligned}$$

For n large enough,

$$\lesssim \frac{\Gamma(M)}{\prod_{j=0}^{K_0} \Gamma(MG_j)} \int_{S_n} (1 - \sum_{j=1}^{K_0} p_j)^{MG_0 - 1} dp_1 \dots dp_{K_0} \times M^{-K_0(d+1)} \delta_n^{K_0 d}$$

We now introduce the following change of variable

$$\begin{cases} p_0 = 1 - \sum_{j=1}^{K_0} p_j \\ p_1 = p_1 \\ \vdots \\ p_{K_0 - 1} = p_{K_0 - 1} \end{cases} \Leftrightarrow \begin{cases} p_{K_0} = 1 - \sum_{j=0}^{K_0 - 1} p_j \\ p_1 = p_1 \\ \vdots \\ p_{K_0 - 1} = p_{K_0 - 1} \end{cases}$$

which Jacobian is equal to 1 so that

$$\lesssim \frac{\Gamma(M)}{\prod_{j=0}^{K_0}\Gamma(MG_j)} \int p_0^{MG_0-1} \mathbb{1}_{(p_0 \leq K_0\delta_n)} \prod_{j=1}^{K_0-1} \mathbb{1}_{(|p_j-\varpi_j^0|\leq\delta_n)} dp_0 \dots dp_{K_0-1} \times M^{-K_0(d+1)} \delta_n^{K_0 d}$$

Where the constraint on p_0 comes from

$$\begin{aligned} |p_{K_0} - \varpi_{K_0}^0| \leq \delta_n &\Rightarrow p_0 \leq 1 - \varpi_{K_0}^0 + \delta_n - \sum_{j=1}^{K_0-1} p_j \\ &\Rightarrow p_0 \leq 1 - \sum_{j=1}^{K_0} \varpi_j^0 + \delta_n + (K_0-1)\delta_n \\ &\Rightarrow p_0 \leq K_0\delta_n. \end{aligned}$$

Hence,

$$\Pi(N_n(P) \leq \delta_n | M) \lesssim \frac{\Gamma(M)}{\prod_{j=0}^{K_0}\Gamma(MG_j)} (K_0\delta_n)^{MG_0} \delta_n^{K_0-1} \times M^{-K_0(d+1)-1} \delta_n^{K_0 d}$$

Integrating with respect to the prior distribution Π_M on M, supported on $[\zeta, \infty]$ for some $\zeta > 0$ yields

$$\begin{aligned} \Pi(N_n(P) \leq \delta_n) &\lesssim \int_\zeta^\infty \frac{\Gamma(M)}{\prod_{j=0}^{K_0}\Gamma(MG_j)} (K_0\delta_n)^{MG_0} \delta_n^{K_0-1} \times M^{-K_0(d+1)-1} \delta_n^{K_0 d} d\Pi_M(M) \\ &\lesssim \delta_n^{K_0-1+dK_0+\zeta G_0} \underbrace{\int_\zeta^\infty \frac{\Gamma(M)}{\prod_{j=0}^{K_0}\Gamma(MG_j)} (K_0)^{MG_0} \times M^{-K_0(d+1)-1} d\Pi_M(M)}_{(*)} \end{aligned}$$

Note that by Stirling's approximation,

$$\frac{\Gamma(M)}{\prod_{j=0}^{K_0}\Gamma(MG_j)} \underset{\infty}{\sim} M^{(K_0-1)/2} \exp\left\{ -M \sum_{j=0}^{K_0} G_j \log(G_j) \right\}.$$

Hence, integral $(*)$ converges if and only if

$$\int_\zeta^\infty M^{(K_0-1)/2 - K_0(d+1) - 1} \exp\left\{ M \left(-\sum_{j=0}^{K_0} G_j \log(G_j) + G_0 \log(K_0) \right) \right\} d\Pi_M(M) < \infty$$

which is guaranteed by Assumption **A5** since $-\sum_{j=0}^{K_0} G_j \log(G_j)$ and G_0 can be made respectively as close to zero and one as desired provided one chooses ϵ small enough.

As required, we conclude that

$$\Pi(N_n(P) \leq \delta_n) \lesssim \delta_n^{K_0 - 1 + dK_0 + \zeta - \varepsilon}$$

where $\varepsilon > 0$ can be chosen as small as desired. □

The results obtained in Theorem 3.1 and subsequently in Corollary 3.2 are of interest when one is able to compute the Bayes Factor comparing a parametric model $\mathfrak{M}_0$ against a nonparametric alternative (here represented by the Dirichlet Process mixture model). Recall that the Bayes Factor, which can be defined as

$$BF_{0,DP} = \frac{m_0(\boldsymbol{y})}{m_{DP}(y)}$$

is rarely available in closed form. Estimating the numerator can be done using the methods prescribed in Chapter 2 when $\mathfrak{M}_0$ is a finite mixture model. However, estimating $m_{DP}(\boldsymbol{y})$ is a non-trivial problem that still lacks a fully satisfactory resolution. To the best of our knowledge, the problem of evidence approximation for the DPM has only been addressed in Basu and Chib 2003. In the following Section, we provide a review of this algorithm and propose an alternative method based on Geyer 1994 Reverse Logisitic Regression solution. We also assess the performance of these algorithms in an empirical study.

3.4 Marginal likelihood estimation for the Dirichlet Process Mixture model

3.4.1 An adaptation of Chib's estimator to the DPM

As extensively discussed in Chapter 2 Section 2.3.2, Chib's algorithm (Chib 1995) is often deemed to be the gold standard of marginal likelihood estimation in finite mixture models. The difficulty of applying Chib's identity (also known as the *candidate's representation*) to the DPM lies in the intractability of the likelihood functions $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ and $p(\boldsymbol{y}|P)$ where P denotes the mixing measure distributed according to $DP(M, G_0)$. Indeed, while the former can only be written in closed form when the allocations $\boldsymbol{z}$ are known, the latter is composed of infinitely many terms. The fundamental idea of Basu and Chib 2003 is to consider instead the integrated likelihood

$$p(\boldsymbol{y}|M) = \int p(\boldsymbol{y}|P) dDP(P; M, G_0) \tag{3.15}$$

and to write

$$m_{DP}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M^0)\pi(M^0)}{\pi(M^0|\boldsymbol{y})}$$

which holds for all $M^0 > 0$.
Unlike its finite mixture counterpart, on top of estimating the posterior distribution that is not available in closed form, Chib's algorithm for the DPM can only be used

if a sensible estimator can be found to estimate the intractable integral defining the integrated likelihood (3.15). We here describe the strategy prescribed by Basu and Chib 2003.

Estimating the posterior ordinate $\pi(M^0|\boldsymbol{y})$. Conveniently, this step can be conducted in a similar fashion as in the finite mixture algorithm by considering the augmented posterior distribution $\pi(M^0|\eta, K_+, \boldsymbol{y})$ and using the integrated estimator

$$\hat{\pi}(M^0|\boldsymbol{y}) = \frac{1}{T_1}\sum_{t=1}^{T_1}\pi(M^0|\eta^{(t)}, K_+^{(t)}, \boldsymbol{y}) \tag{3.16}$$

where $\pi(M^0|\eta^{(t)}, K_+^{(t)}, \boldsymbol{y})$ is the mixture of two Gamma distributions defined in (3.14) and $\{\eta^{(t)}, K_+^{(t)}\}_{t=1}^{T_1} \sim \pi(\eta, K_+|\boldsymbol{y})$ are directly available from the output of the Gibbs sampler presented in Algorithm 11.

Estimating the likelihood ordinate $p(\boldsymbol{y}|M^0)$. Estimating the likelihood (3.15) is less straightforward and cannot be done by a simple post-processing step of the Gibbs output. Instead, Basu and Chib 2003 suggests to adopt a Sequential Importance Sampling strategy of data imputation, following the seminal article by Kong et al. 1994. The framework is the same as the one prescribed for the SIS estimator derived in Chapter 2 Section 2.4.2. The core idea is still to define an approximation to the posterior distribution in which the latent variable $\boldsymbol{z}$ is imputed sequentially

$$\pi^*(z_1, \ldots, z_n|M^0, \boldsymbol{y}) = \pi(z_1|M^0, y_1)\prod_{i=2}^{n}\pi(z_i|M^0, \boldsymbol{y}_{1:i}, \boldsymbol{z}_{1:i-1}). \tag{3.17}$$

Then one can notice that

$$\begin{aligned}
\pi(\boldsymbol{z}|M^0, \boldsymbol{y}) \times \frac{1}{\pi^*(\boldsymbol{z}|M^0, \boldsymbol{y})} &= \frac{p(\boldsymbol{z}, \boldsymbol{y}|M^0)}{p(\boldsymbol{y}|M^0)} \\
&\quad \times \left(\frac{p(y_1|M^0)}{p(z_1, y_1|M^0)}\frac{p(z_1, \boldsymbol{y}_{1:2}|M^0)}{p(\boldsymbol{z}_{1:2}, \boldsymbol{y}_{1:2}|M^0)} \cdots \frac{p(\boldsymbol{z}_{1:n-1}, \boldsymbol{y}|M^0)}{p(\boldsymbol{z}, \boldsymbol{y}|M^0)}\right) \\
&= \frac{p(y_1|M^0)\prod_{i=2}^{n} p(y_i|M^0, \boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1})}{p(\boldsymbol{y}|M^0)} \\
&= \frac{w(\boldsymbol{z}, \boldsymbol{y})}{p(\boldsymbol{y}|M^0)}
\end{aligned}$$

where $w(\boldsymbol{z}, \boldsymbol{y}, M^0) := p(y_1|M^0)\prod_{i=2}^{n} p(y_i|M^0, \boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1})$.
This leads to the following identity

$$\begin{aligned}
\int w(\boldsymbol{z}, \boldsymbol{y}, M^0)\pi^*(\boldsymbol{z}|M^0, \boldsymbol{y})d\boldsymbol{z} &= \int p(\boldsymbol{y}|M^0)\pi(\boldsymbol{z}|M^0, \boldsymbol{y})d\boldsymbol{z} \\
&= p(\boldsymbol{y}|M^0)
\end{aligned}$$

which induces the unbiased estimator of the integrated likelihood

$$\hat{p}(\boldsymbol{y}|M^0) = \frac{1}{T}\sum_{t=1}^{T} w(\boldsymbol{z}^{(t)}, \boldsymbol{y}, M^0)$$

for a sample $\{\boldsymbol{z}^{(t)}\}_{t=1}^{T_2}$ from $\pi^*(\boldsymbol{z}|M^0, \boldsymbol{y})$.

Sampling from $\pi^*(\boldsymbol{z}|M^0, \boldsymbol{y})$ is straightforward when considering the expression for $\Pi(z_i|M, \boldsymbol{z}_{-i}, \boldsymbol{y})$ given in (3.12). It writes

$$\Pi(z_i = k|\boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i}, M) \propto \begin{cases} M\int_\Theta f(y_i|\theta)G_0(d\theta) & \text{if } k = K_+^{i-1}+1 \\ N_k^{i-1}\frac{m_k(\boldsymbol{z}_{1:i-1}\cup\{z_i=k\})}{m_k(\boldsymbol{z}_{i-1})} & \text{if } k = 1,\ldots,K_+^{i-1} \end{cases} \tag{3.18}$$

where K_+^{i-1}, N_k^{i-1} denote respectively the number of non-empty clusters and the number of observations allocated to group $k < K_+^{i-1}$ induced by vector $\boldsymbol{z}_{1:i-1}$.
In a similar way, we can derive

$$\begin{aligned} p(y_i|M, \boldsymbol{z}_{1:i-1}, \boldsymbol{y}_{1:i-1}) = &\sum_{k=1}^{K_+^{i-1}} \frac{N_k^{i-1}}{M+i-1}\frac{m_k(\boldsymbol{z}_{1:i-1}\cup\{z_i=k\})}{m_k(\boldsymbol{z}_{i-1})} \\ &+ \frac{M}{M+i-1}\int_\Theta f(y_i|\theta)G_0(d\theta) \end{aligned} \tag{3.19}$$

which yields the strategy described in Algorithm 12 below.

Algorithm 12 : Basu and Chib 2003

```
Sample {M^(t), η^(t)}_{t=1}^{T_1} from the posterior π(M, η|y) using Algorithm 11;
Set M^0 = (1/T_1) Σ_t M^(t) or any other point with high posterior probability.;
Compute π̂(M^0|y) using (3.16);
for t = 1, ..., T_2 do
    for i = 1, ..., n do
        Compute u_i^(t) = p(y_i|M^0, z_{1:i-1}, y_{1:i-1}) using (3.19);
        Re-use the computations of the step above to sample z_i from
          Π(z_i = k|z_{1:i-1}, y_{1:i}, M^0) using (3.18);
    end
end
Compute p̂(y|M^0) = (1/T_2) Σ_{t=1}^{T_2} Π_{i=1}^{n} u_i^(t);
Return m_Chib(y) = p̂(y|M^0)π(M^0)/π̂(M^0|y)
```

3.4.2 A novel approach based on Reverse Logistic Regression

Introduced by Geyer 1994, Reverse Logistic Regression (RLR) is a biased Importance Sampling approach to the issue of estimating normalizing constants. Let $\pi_1(\theta)$ and $\pi_2(\theta)$ denote two distributions, with $\pi_1 = \tilde{\pi}_1/c_1$ only determined up to a proportionality constant where $c_1 := \int \tilde{\pi}_1(\theta)d\theta$. If one has access to samples

$\{\theta_1^{(t)}\}_{t=1}^{T_1}$ and $\{\theta_2^{(t)}\}_{t=1}^{T_2}$ respectively from π_1 and π_2, the trick suggested by Geyer 1994 is to think of the samples as arising from the two component mixture

$$h_{mix}(\theta) = \frac{T_1}{|T|}\tilde{\pi}_1(\theta)/c_1 + \frac{T_2}{|T|}\pi_2(\theta)$$

where $|T| := T_1 + T_2$.
This representation is completely artificial and in fact counterintuitive since we know exactly from which component each sample is generated. However, it allows for a new approach to the problem of estimating c_1. Indeed, consider the probability p_1 that some θ belongs to group 1,

$$p_1(\theta, c_1) = \frac{(T_1/|T|)\tilde{\pi}_1(\theta)/c_1}{(T_i/|T|)\tilde{\pi}_1(\theta)/c_1 + (T_2/|T|)\pi_2(\theta)}$$

then estimating c_1 can be done by maximizing the log quasi-likelihood

$$\ell(c_1) = \sum_{t=1}^{T_1} \log p_1(\theta_1^{(t)}, c_1) + \sum_{t=1}^{T_2} \log(1 - p_1(\theta_2^{(t)}, c_1)). \tag{3.20}$$

This in fact defines a logistic regression inference scheme, except for the fact that the *response* variable happens to be known while the regressors θ are random, hence the name Reverse Logistic Regression. Notice that it is crucial that π_2 be a normalized distribution for identifiability reasons. Indeed, the likelihood $\ell(\theta, c_1, c_2)$ is equal to $\ell(\theta, \kappa c_1, \kappa c_2)$ for any constant κ, which makes the maximum likelihood estimator for c_1 and c_2 unique up to a multiplicative constant.

Chen and Shao 1997 show that this procedure is in fact essentially equivalent to Bridge Sampling, as defined in Chapter 2 Section 2.3.3. As such, standard results of Importance sampling apply and in particular the choice of the importance distribution π_2 is crucial for a good estimation of c_1. To better understand the connection to Bridge Sampling, notice that, assuming for simplification that $T_1 = T_2 = T$,

$$\begin{aligned}\frac{\partial \ell(c_1)}{\partial c_1} &= \sum_{t=1}^{T} \frac{\partial p_1/\partial c_1}{p_1(\theta_1^{(t)})} - \sum_{t=1}^{T} \frac{\partial p_1/\partial c_1}{1 - p_1(\theta_2^{(t)}, c_1)} \\ &= -\frac{1}{c_1}\sum_{t=1}^{T} \frac{\pi_2(\theta_1^{(t)})}{\tilde{\pi}_1(\theta_1^{(t)})/c_1 + \pi_2(\theta_1^{(t)})} + \frac{1}{c_1}\sum_{t=1}^{T} \frac{\tilde{\pi}_1(\theta_1^{(t)})/c_1}{\tilde{\pi}_1(\theta_1^{(t)})/c_1 + \pi_2(\theta_1^{(t)})}\end{aligned}$$

so that the maximum likelihood estimator $\widehat{c_1}_T$ verifies

$$\sum_{t=1}^{T} \frac{\pi_2(\theta_1^{(t)})}{\tilde{\pi}_1(\theta_1^{(t)})/\widehat{c_1}_T + \pi_2(\theta_1^{(t)})} \Big/ \sum_{t=1}^{T} \frac{\tilde{\pi}_1(\theta_2^{(t)})/\widehat{c_1}_T}{\tilde{\pi}_1(\theta_2^{(t)})/\widehat{c_1}_T + \pi_2(\theta_2^{(t)})} = 1.$$

Letting T tend to infinity, and denoting the limit of $\widehat{c}_{1T}$ by $\bar{c}_1$

$$\begin{aligned}
&\int \frac{\pi_2(\theta)\pi_1(\theta)}{\tilde{\pi}_1(\theta)/\bar{c}_1 + \pi_2(\theta)} d\theta \Big/ \int \frac{\tilde{\pi}_1(\theta)/\bar{c}_1\pi_2(\theta)}{\tilde{\pi}_1(\theta)/\bar{c}_1 + \pi_2(\theta)} d\theta = 1 \\
\Leftrightarrow &\int \frac{\pi_2(\theta)\tilde{\pi}_1(\theta)/\bar{c}_1}{\tilde{\pi}_1(\theta)/\bar{c}_1 + \pi_2(\theta)} d\theta \Big/ \int \frac{\tilde{\pi}_1(\theta)/\bar{c}_1\pi_2(\theta)}{\tilde{\pi}_1(\theta)/\bar{c}_1 + \pi_2(\theta)} d\theta = c_1/\bar{c}_1 \\
\Leftrightarrow &\bar{c}_1 = c_1
\end{aligned} \tag{3.21}$$

under some regularity assumption.
Hence, $\widehat{c}_{1T} \longrightarrow c_1$ and identifying equation (3.21) with the Bridge Sampling identity (2.23) shows that RLR is asymptotically equivalent to optimal Bridge Sampling. Chen and Shao 1997 provide an asymptotic estimate of the relative error of $\widehat{c}_{1T}$ w.r.t c_1 given by

$$\mathbb{E}\left[(\widehat{c}_{1T} - c_1)^2/c_1^2\right] \approx \frac{4}{T}\left[\left\{\int \frac{\pi_1(\theta)\pi_2(\theta)}{0.5\pi_1(\theta) + 0.5\pi_2(\theta)} d\theta\right\}^{-1} - 1\right]$$

which stresses the need for an importance distribution π_2 with as much overlap with π_1 as possible with minimum requirement that $\int \pi_1(\theta)\pi_2(\theta)d\theta > 0$.

Returning to the issue of estimating the marginal likelihood $m_{DP}(\boldsymbol{y})$, the application of RLR with $\tilde{\pi}_1(\boldsymbol{z}, M) := p(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}|M)\pi(M)$ boils down to finding a suitable importance distribution that is a good approximation to the posterior $\pi(\boldsymbol{y}|\boldsymbol{z}, M)$. We suggest using

$$\pi_2(\boldsymbol{z}, M) := \pi^*(\boldsymbol{z}|\boldsymbol{y}, M)\pi(M)$$

where $\pi^*(\boldsymbol{z}|\boldsymbol{y}, M)$ is the sequential approximation to the posterior distribution defined in (3.17). Notice that $\int \pi_2(\boldsymbol{z}, M)d\boldsymbol{z}dM = 1$ and that sampling from π_2 can be achieved by first sampling M from the prior, before sampling sequentially the allocations $\boldsymbol{z}$ using equation (3.18). We summarize the procedure in Algorithm 13 below.

Algorithm 13 : Reverse Logistic Regression for marginal likelihood estimation in a DPM (RLR-SIS).

1 Sample $\{\boldsymbol{z}_1^{(t)}, M_1^{(t)}\}_{t=1}^{T_1}$ from the posterior $\pi(M, \boldsymbol{z}|\boldsymbol{y})$ using Algorithm 11;
2 **for** $t = 1, \ldots, T_2$ **do**
3 | Sample $M_2^{(t)}$ from the prior;
4 | Sample $\boldsymbol{z}_2^{(t)}|M_2^{(t)}, \boldsymbol{y}$ using (3.18);
5 **end**
6 Optimize the log quasi-likelihood $\ell(c_1)$ (3.20);
7 Return $m_{DP}(\boldsymbol{y}) = \arg\max_{c_1} \ell(c_1)$

3.5 Simulation study

Our ambition for this section is twofold. First, we assess and compare the algorithm of Basu and Chib 2003 to our proposed estimator as well as other alternatives which, to our knowledge, has never been done before. Second, we provide an empirical illustration to the Bayes Factor consistency result obtained in Corollary 3.2.

Unless specified otherwise, we fit Normal Dirichlet Process mixture models and use the *conditionally conjugate* Normal-Inverse Gamma prior G_0 for the location and scale parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ as described in Example 1 and equation (2.10) of Chapter 2 which is defined for all $k = 1, \dots, K$ by $\sigma_k^2 \sim \Gamma^{-1}(a, b)$ and $\mu_k | \sigma_k^2 \sim \mathcal{N}(\mu_0, \sigma_k^2/\lambda)$ where Γ^{-1} is the inverse gamma distribution in the shape and scale parametrization. The hyperparameters (a, b, μ, λ) are derived empirically following recommendations from Raftery 1996 : $a = 1.28, b = 0.36(\overline{y^2} - \overline{y}^2), \mu_0 = \bar{y}, 1/\lambda = (y_{max} - y_{min})/2.6$. Finally the prior on the concentration parameter M is chosen to be a Gamma distribution with shape and scale $(1, 1)$, as commonly done by practitioners. Note that this choice of prior ensures that $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z}) := \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2|\boldsymbol{y}, \boldsymbol{z})$ is available in closed form, and that it makes the Gibbs sampling strategy given in Algorithm 11 readily applicable.

3.5.1 Experiment 1 : Galaxies data.

In this section, we evaluate the relative performances of Basu and Chib 2003's algorithm with other competing alternatives. To do so, we once again start by using the `galaxies` data that contains the radial velocity of 82 galaxies and consider a Dirichlet Process mixture of normal distributions with unknown location and scale parameters.

Note that, while for finite mixtures, model complexity is a function of the number of mixture components K, most difficulties arise in the DPM for a growing number of observations n. This is partly due to the fact that the number of non-empty clusters K_+ is $\mathcal{O}(M \log(n))$. Hence we design three different scenarios with subsets of the galaxies data of size respectively 6, 36 and 82, as shown on Figure 3.3. One can note in particular that the Arithmetic Mean estimator (AME, defined in 2.3.1 of Chapter 2), which shows rather good results for a small number of observations, fails to converge as soon as the amount of data becomes moderately large. Since we are not able to make the AME converge (i.e to make its Monte Carlo variance low) for even a prohibitively high computational time, we use Reverse Logistic Regression where the prior is used as the importance distribution (hereafter RLR-Prior) as a reference value. This method is indeed independent of both Basu and Chib 2003's algorithm and RLR with a SIS adversarial distribution since it does not depend on the sequential imputation scheme (3.18). It is interesting to note that RLR-Prior yields rather satisfactory results despite such a naïve choice of importance distribution. For non-conjugate Dirichlet Process Mixture models, this rather inexpensive estimator could be a good, easy-to-implement, first approach to the marginal likelihood estimation problem.

It is interesting to see that both Chib's algorithm and our candidate RLR-SIS yield accurate estimates of the marginal likelihood for all scenarios considered. The

Harmonic Mean estimator (HME, cf Section 2.3.1 of Chapter 2), however, seems to suffer from a simulation pseudo-bias as also discussed in Section 2.3.1 of Chapter 2, which leads to an overestimation of the marginal likelihood. Note that this phenomenon is probably combined to an explosive variance which is typical of the HME, as already pointed by Radford M. Neal in his discussion of Newton and Raftery 1994.

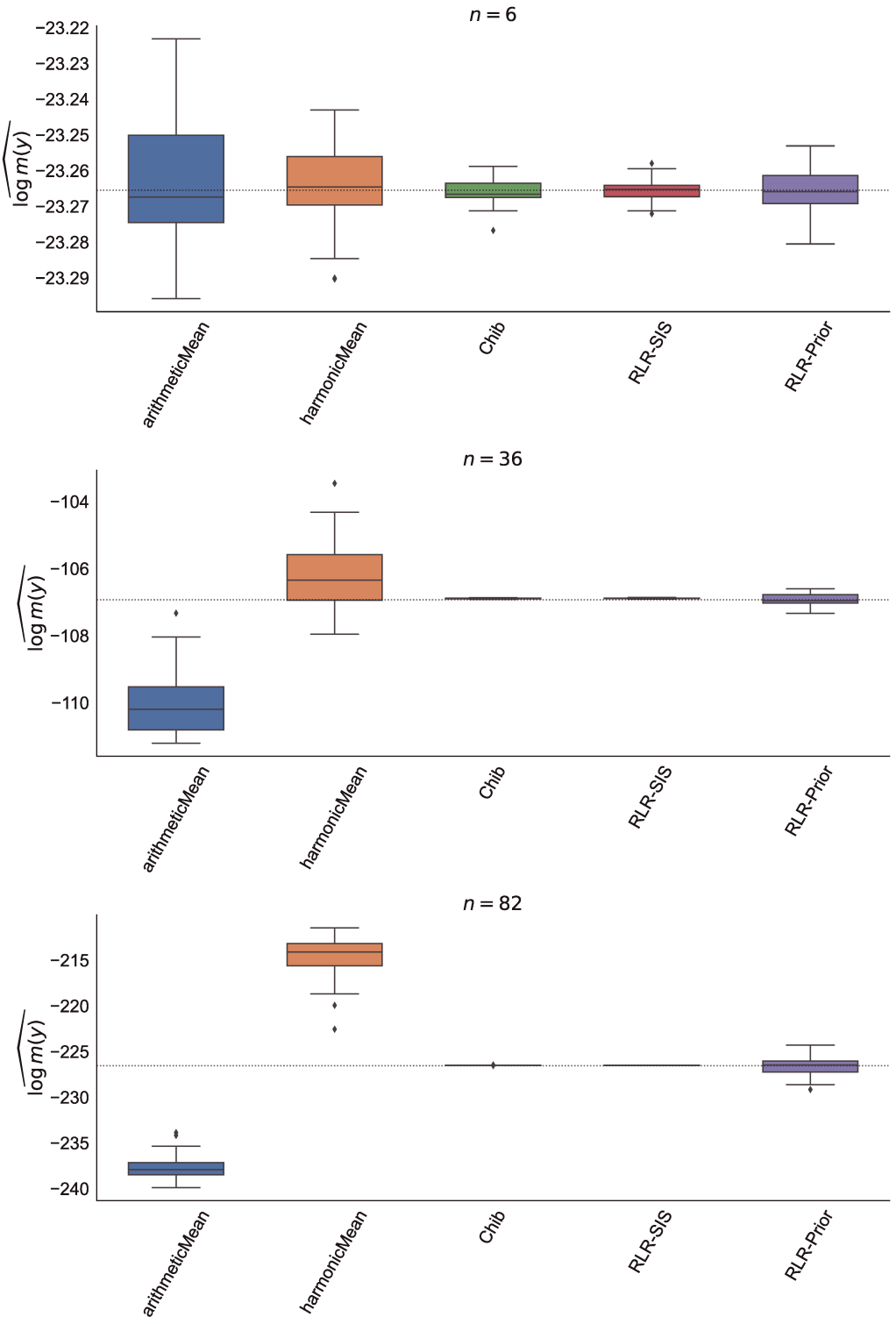


Figure 3.3: `galaxies` data. Boxplots of the different methods (20 repetitions each) for different values of n. The dashed line corresponds to the mean of the RLR-prior estimator. The choice of tuning parameters is given in Table 3.1 in Appendix.

For the full `galaxies` data, Figure 3.4 displays the decrease of the MSE as a function of time and shows that for a given allocated time, the error produced by

Basu's and Chib's estimator is greater than that of RLR-SIS by about a factor $\exp(0.3) \approx 1.35$.

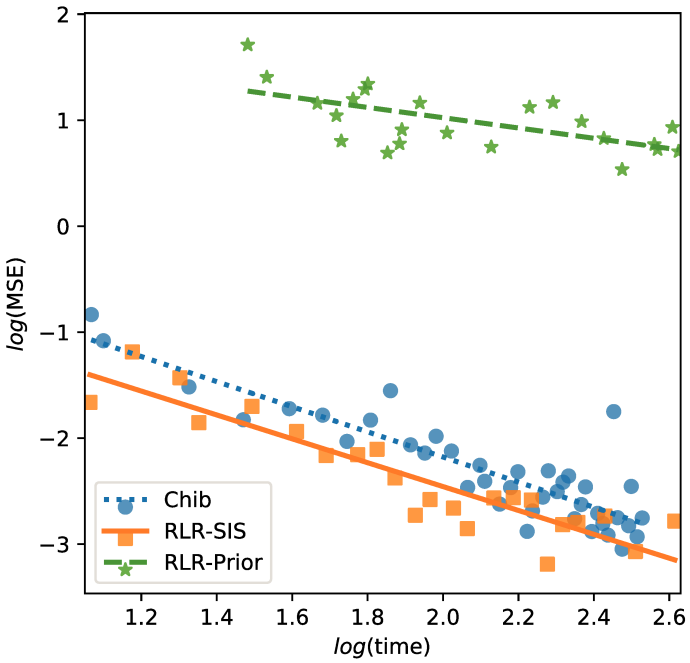


Figure 3.4: `galaxies` data. MSE vs time. The dashed lines represent the Ordinary Least Square line fitting fitting the log of the MSE and the log of the time for each of the different methods.

3.5.2 Experiment 2 : Synthetic data, $n = 1000$.

We now assess the scalability of both methods on a synthetic data set of 1000 observations, arising from a 6-component mixture of normal distributions with weights $\boldsymbol{\varpi} = (0.2, 0.01, 0.27, 0.19, 0.19, 0.14)$, means $\boldsymbol{\mu} = (2.5, -6.2, -5.3, -4.5, 2.8, 11.55)$ and scales $\boldsymbol{\sigma} = (2, 2, 2, 2, 2, 2)$. Although we cannot compute analytically the true value of the marginal likelihood, we do know that both Chib's algorithm and RLR-SIS are consistent, as illustrated in the previous experiment. That is, provided we do not observe a large simulation-related variance, the estimator can be trusted. The tuning parameters used for the experiment, as well as the run times are given in the caption of Figure 3.5. The run time corresponds to the total time required to compute 20 repetitions of each estimator on 10 cores, each core being an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz. Note that no parallelization of the embarrassingly parallel SIS step was done (which would decrease the total run time of both algorithms dramatically, but by the same factor). On Figure 3.5, it is clear that Chib's estimator suffers from a large variance, which in turn translates into a downward bias on the *log*-scale. On the other hand, RLR-SIS exhibits much more stability. Note that although more iterations and computational time are allocated to Chib's estimator, this does not seem to be enough to correct Chib's higher variance while RLR-SIS

yields a more accurate estimate of the marginal likelihood within a much lower run time.

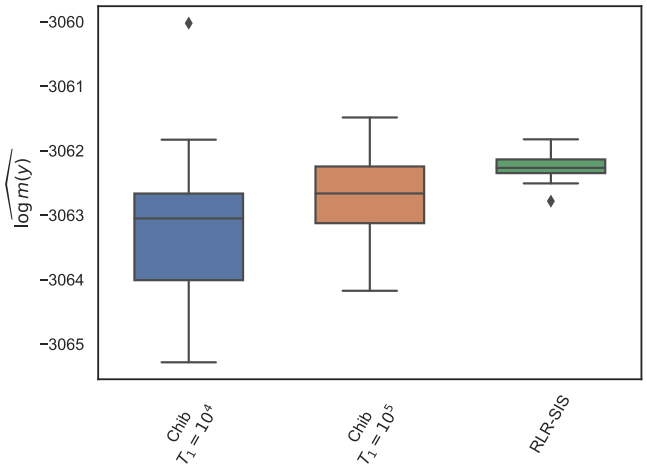


Figure 3.5: Boxplots of the marginal likelihoods estimates, 20 repetitions each. Dirichlet process mixture model, synthetic data, $n = 1000$. Chib left hand side : $T_1 = 10^4$, $burnIn = 10^3$, $T_2 = 600$, Run time : 04:09:51. Chib right hand side $T_1 = 10^5$, $burnIn = 10^4$, $T_2 = 2000$, Run time : 34:01:18. RLR-SIS : $T_1 = 10^4$, $burnIn = 10^3$, $T_2 = 600$, Run time : 06:23:12.

3.5.3 Experiment 3 : Testing a finite mixture against a DPM.

In this experiment, we check whether the Bayes factor converges to infinity under the null hypothesis that data arises from a finite mixture with K_0 components, when fitting a DPM. That is, we want to check whether $BF_{K,DP} := \frac{m_K(\boldsymbol{y})}{m_{DP}(\boldsymbol{y})} \underset{n\to\infty}{\longrightarrow} \infty$, for K not necessarily equal to K_0. Note that we here relax some of the assumptions of Theorem 3.1 and Corollary 3.2. In particular we use the same prior as described above which implies that M has support on $(0, \infty)$ a priori, and we consider location-scale mixtures of Normal distributions, which are not covered by our consistency result. The goal of this experiment is to show empirically that such a result still holds for more general mixtures.

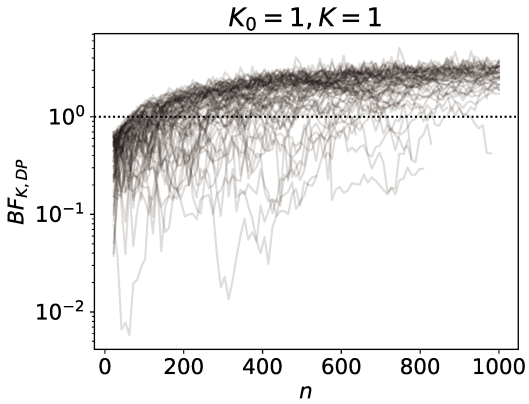
To do so, 100 data sets $\boldsymbol{y}$ are generated from a Data Generating Process (DGP) P_0, where P_0 is a finite mixture. The Bayes Factor is then computed for successive values of n and the resulting 'Bayes Factor paths' are displayed on Figure 3.6. We consider three different DGPs P_0 given by

$$\begin{cases} P_0^{(1)} = \mathcal{N}(0, 4) \\ P_0^{(2)} = 0.3\mathcal{N}(-3, 1) + 0.2\mathcal{N}(4, 1) + 0.5\mathcal{N}(12, 1) \\ P_0^{(3)} = 0.3\mathcal{N}(0, 1) + 0.2\mathcal{N}(1, 1) + 0.5\mathcal{N}(2, 1). \end{cases}$$

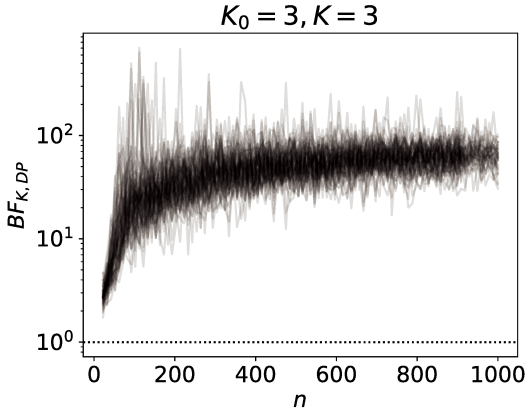
In Figure 3.6a, we fit a one component mixture of Normal distributions, or in other words a Normal distribution, on 100 data sets generated from $P_0^{(1)}$. Observe that

most Bayes Factor paths are above the dashed line drawn at $y = 1$. Empirically, we can see that the Bayes Factor appears to point towards the finite mixture, as it seems that $P_0^{(1)}(BF_{K,DP} > 1) \longrightarrow 1$. Figure 3.6b shows the Bayes Factor paths of a $3-$component mixture against the DPM for data arising from $P_0^{(2)}$. It is interesting to notice that in this scenario the Bayes Factor seems to converge much faster than in the previous case. This is slightly counter-intuitive since we expect the DPM to yield a better fit on multimodal data which should induce a slower convergence of the Bayes Factor for $K_0 = 3$ than for $K_0 = 1$. However, this could be due to the rather well-separated modal configuration of data generated from $P_0^{(2)}$ which should lead to a very good fit for the finite mixture model with 3 components. This intuition is confirmed in Figure 3.6c where the DGP used, $P_0^{(3)}$ still has 3 components, but much less separated. We observe indeed a much slower convergence of the Bayes Factor to infinity, although most Bayes Factor paths seem to cross the reference line $y = 1$ for increasing values of n. On Figure 3.6d, DGP $P_0^{(3)}$ is still used but a 5-component Normal mixture model is fitted instead. In this scenario, the DPM alternative seems to yield a rather good fit, although most Bayes Factor paths seem to favour the finite mixture as n increases.

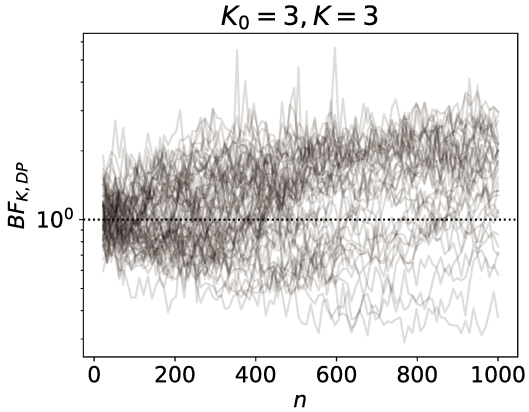
We also consider an under-specified scenario in which $K_0 = 3$ (using $P_0^{(2)}$ as the DGP) while $K = 1$ (Figure 3.6e) and $K = 2$ (Figure 3.6f). This relates to the case $(i.i.i)$ of Corollary 3.2. We clearly see that empirically the Bayes Factor converges to 0, indicating a better fit of the DPM. Subsequently, an interesting test for checking the validity of a finite mixture model in a given population is to consider a mixture with a conservatively large number of components K. If the Bayes Factor against the DPM points towards the finite mixture, it is a good indication that a finite mixture with $K_0 \leq K$ components is a good fit for the data. Notice that this relates to the strategy of overfitting mixtures suggested by the results Rousseau and Mengersen 2011.



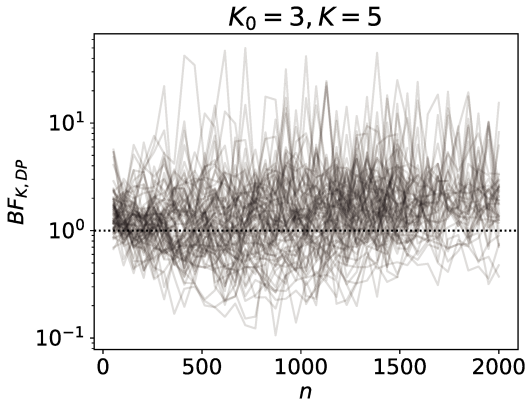
(a) DGP : $P_0^{(1)}$. Bayes Factor paths for 100 data sets from $P_0^{(1)}$ of a 1-component mixture against the DPM.



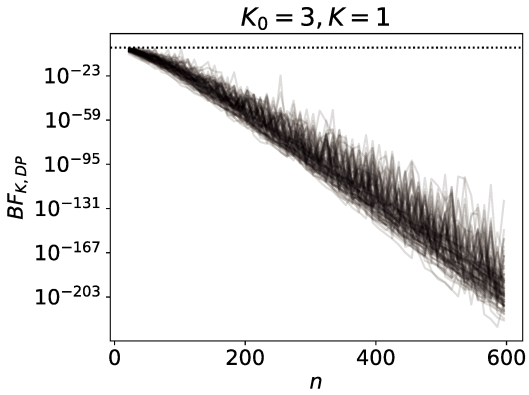
(b) DGP : $P_0^{(2)}$. Bayes Factor paths for 100 data sets from $P_0^{(2)}$ of a 3-component mixture against the DPM.



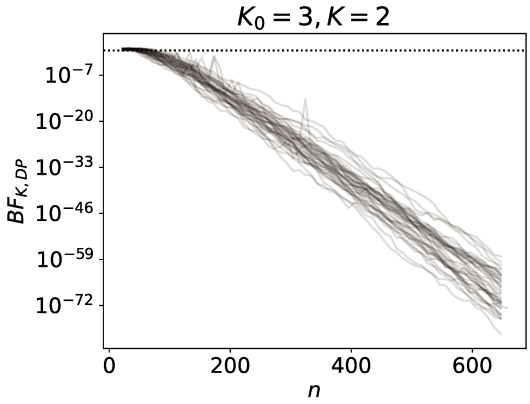
(c) DGP : $P_0^{(3)}$. Bayes Factor paths for 100 data sets from $P_0^{(3)}$ of a 3-component mixture against the DPM.



(d) DGP : $P_0^{(3)}$. Bayes Factor paths for 100 data sets from $P_0^{(3)}$ of a 5-component mixture against the DPM.



(e) DGP : $P_0^{(2)}$. Bayes Factor paths for 100 data sets from $P_0^{(2)}$ of a 1-component mixture against the DPM.



(f) DGP : $P_0^{(2)}$. Bayes Factor paths for 100 data sets from $P_0^{(2)}$ of a 2-component mixture against the DPM.

Figure 3.6: Experiment 3. Bayes Factor paths for different scenarios. The value of K in the title of each plot corresponds to the order of the finite mixture model fitted on the data, that is compared to the DPM through the Bayes Factor. The dashed line is $y = 1$. The y-axis is in log scale. Notice the differences in the x-axes.

3.6 Conclusion and perspectives

In this chapter, we proved in Theorem 3.1 and Corollary 3.2 the consistency of the Bayes Factor associated to the problem of testing a parametric model against a Dirichlet Process Mixture model, both under the null and the alternative, which was an open problem so far. In addition, Lemma 3.3 provides a solution to the unaddressed challenge of upper bounding the prior mass of L_1-neighborhoods of the true mixture density f_0 under the DP prior. An immediate application is goodness-of-fit tests that compare a parametric null to a nonparametric alternative.

However, we noticed that there does not seem to exist any reference method widely used by practitioners for numerically estimating the marginal likelihood of a Dirichlet Process Mixture model. We benchmarked Basu and Chib 2003's approach with other algorithms, including a method based on Geyer 1994's reverse logistic regression that is reliable and scales better with the number of observations n. Our simulation study supports the use of Geyer's RLR method with a SIS importance distribution as a way to estimate the marginal likelihood of a DPM.

An interesting research avenue would be to identify scalable evidence estimation techniques for non-conjugate Dirichlet Process mixtures. The Reverse Logistic Regression method seems like a good framework to achieve this goal, provided one can derive a good adversarial distribution π^* for such models. As to the asymptotics of the marginal likelihood under a DPM, it would be interesting to extend Theorem 3.1 to location-scale Dirichlet Process mixtures. Our computational simulations support that such a result is still valid for this wider class of models.

3.A Appendix

3.A.1 Technical lemmas

The following lemma by Yamato 1984 provides an expression for the characteristic function ψ of $\mu(P) := \int \theta dP(\theta)$ for P a realization of the Dirichlet Process $DP(M, G_0)$. Lemma 3.5 subsequently establishes that ψ is integrable. This is at the core of the proof of Lemma 3.3 which makes use of the existence of a continuous density function for the random variable $\mu(P)$.

Lemma 3.4. *(Yamato 1984) Under Assumption A4, the characteristic function ψ of $\mu(P)$ can be written as*

$$\psi(\boldsymbol{z}) = \mathbb{E}\left[\prod_{j=1}^{\infty} \varphi(\varpi_j \boldsymbol{z})\right],\ \boldsymbol{z} \in \mathbb{R}^d$$

Lemma 3.5. *For all $d \geq 1$, if $P = \sum_{j=1}^{\infty} \varpi_j \delta_{\theta_j} \sim DP(M, G_0)$,*

$$\mathbb{E}_P\left[\left(\sum_{j=1}^{d+1} \varpi_j^2\right)^{-\frac{d}{2}} \middle| M\right] < \infty$$

Proof. We first notice that

$$\begin{cases} V_1 = \varpi_1 \\ V_2(1-V_1) = \varpi_2 \\ \vdots \\ V_{d+1}(1-V_d)\dots(1-V_1) = \varpi_{d+1} \end{cases} \Leftrightarrow \begin{cases} V_1 = \varpi_1 \\ V_2 = \frac{\varpi_2}{1-\varpi_1} \\ \vdots \\ V_{d+1} = \frac{\varpi_{d+1}}{1-\sum_i^d \varpi_i} \end{cases}$$

where $V_i \overset{i.i.d}{\sim} Beta(1, M)$. Hence,

$$\begin{aligned} I :=& \mathbb{E}\left[\frac{1}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}} \middle| M\right] \\ =& \left(\frac{\Gamma(M)}{\Gamma(1+M)}\right)^{d+1} \int_{[0,1]^{d+1}} \frac{(1-V_1)^{M-1}\dots(1-V_{d+1})^{M-1}}{(V_1^2 + \dots + V_{d+1}^2(1-V_d)^2\dots(1-V_1)^2)^{d/2}} dV_1 \dots dV_{d+1} \end{aligned}$$

Let $B_\varepsilon = \{(V_1, \dots, V_{d+1}) : \exists i \in \{1, \dots, d+1\},\ \varpi_i > \varepsilon\}$ and fix $\varepsilon = 1/(2(d+1))$. Then,

$$\begin{aligned} I \leq& \left(\int_{[0,1]^{d+1} \cap B_\varepsilon^c} \frac{(1-V_1)^{M-1}\dots(1-V_{d+1})^{M-1}}{(V_1^2 + \dots + V_{d+1}^2(1-V_d)^2\dots(1-V_1)^2)^{d/2}} dV_1 \dots dV_{d+1} + \varepsilon^{-d}\right) \\ & \times \left(\frac{\Gamma(M)}{\Gamma(1+M)}\right)^{d+1} \end{aligned}$$

We then use the change of variable given by (3.A.1) which Jacobian is

$$\det(J) = \frac{1}{1-\varpi_1} \times \frac{1}{1-\varpi_1-\varpi_2} \times \cdots \times \frac{1}{1-\sum_i^d \varpi_i}$$

Hence,

$$I \leq \left(\varepsilon^{-d} + \int_{\Delta_{d+1} \cap B_\varepsilon^c} \frac{1}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}} \times \frac{\left(1-\sum_{j=1}^{d+1} \varpi_j\right)^{M-1}}{(1-\varpi_1)\dots(1-\sum_{j=1}^{d} \varpi_j)} d\varpi_1 \dots d\varpi_{d+1} \right) \times \left(\frac{\Gamma(M)}{\Gamma(1+M)} \right)^{d+1}$$

where $\Delta_{d+1} = \{\boldsymbol{\varpi} \in \mathbb{R}^{d+1} : 0 \leq \varpi_j \leq 1, \forall j \text{ and } \sum_j \varpi_j \leq 1\}$.

$$I \leq \left(\frac{\Gamma(M)}{\Gamma(1+M)} \right)^{d+1} \left(\varepsilon^{-d} + 2^d \int_{\Delta_{d+1} \cap B_\varepsilon^C} \frac{\left(1-\sum_{j=1}^{d+1} \varpi_j\right)^{M-1}}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}} d\varpi_1 \dots d\varpi_{d+1} \right)$$

If $M \geq 1$

$$I \leq \left(\frac{\Gamma(M)}{\Gamma(1+M)} \right)^{d+1} \left(\varepsilon^{-d} + 2^d \int_{\Delta_{d+1}} \frac{1}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}} d\varpi_1 \dots d\varpi_{d+1} \right)$$

If $M < 1$

$$I \leq \left(\frac{\Gamma(M)}{\Gamma(1+M)} \right)^{d+1} \left(\varepsilon^{-d} + 2^{d+1-M} \int_{\Delta_{d+1}} \frac{1}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}} d\varpi_1 \dots d\varpi_{d+1} \right)$$

Let $J = \int_{\Delta_{d+1}} 1/(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}} d\varpi_1 \dots d\varpi_{d+1}$. We shall prove that J is finite.

Using the hyperspherical change of variable

$$\begin{cases} \varpi_1 = r\cos\theta_1 \\ \varpi_2 = r\sin\theta_1\cos\theta_2 \\ \vdots \\ \varpi_d = r\sin\theta_1 \dots \cos\theta_d \\ \varpi_{d+1} = r\sin\theta_1 \dots \sin\theta_d \end{cases}$$

where $\theta_1, \dots, \theta_{d-1} \in [0, \pi]$ and $\theta_d \in [0, 2\pi)$. The Jacobian of this transformation is bounded by r^d, and noticing that $r = \left(\sum_{j=1}^{d+1} \varpi_j^2\right)^{\frac{1}{2}}$, we get

$$J \leq \int_0^1 \int_0^{2\pi} \dots \int_0^{\pi} \frac{r^d}{r^d} d\theta_1 \dots d\theta_d dr$$

Hence,

$$J \leq 2\pi \left(\frac{\pi}{2}\right)^{d-1}$$

and

$$\mathbb{E}\left[\frac{1}{(\sum_{j=1}^{d+1} \varpi_j^2)^{\frac{d}{2}}}\middle| M\right] < \infty \quad \square$$

$\square$

Lemma 3.6. *If $P = \sum_{j=1}^{\infty} \varpi_j \delta_{\theta_j} \sim DP(M, G_0)$ where G_0 has a density with respect to Lebesgue measure and verifies*

$$\int |\varphi(u)| du = C_{G_0} < \infty,$$

with φ the characteristic function of G_0 and C_{G_0} a constant, then under Assumption A4 $\mu(P)$ has a bounded density with respect to the Lebesgue measure.

Proof. Using Lemma 3.4, the characteristic function ψ of $\mu(P)$ is equal to

$$\psi(\boldsymbol{z}) = \mathbb{E} \prod_{j=1}^{\infty} \varphi(\varpi_j \boldsymbol{z}),\ P = \sum_j \varpi_j \delta_{\theta_j}$$

where φ is the characteristic function of G_0. Let $\varpi_{(1)} = \max_j \varpi_j$ and note that, in particular, $|\psi(\boldsymbol{z})| \leq \mathbb{E}|\varphi(\varpi_{(1)}\boldsymbol{z})|$. Hence,

$$\int |\psi(\boldsymbol{z})| d\boldsymbol{z} \leq \mathbb{E}\left[\frac{1}{\varpi_{(1)}^d} \int |\varphi(u)| du\right] \leq C_{G_0} \mathbb{E}\left[\left(\frac{\sum_{j=1}^r \varpi_j^2}{r}\right)^{-d/2}\right], \text{ for some } r > 0$$

Using Lemma 3.5, with $r = d + 1$, we obtain,

$$\int |\psi(\boldsymbol{z})| d\boldsymbol{z} \lesssim C_{G_0} \left(\frac{\Gamma(M)}{\Gamma(1+M)}\right)^{d+1}$$

$\square$

Lemma 3.7. *Let $P = \sum_{i=1}^{\infty} \varpi_i \delta_{\theta_i}$ be a realization of the Dirichlet Process. Define*

$$\begin{aligned}\Delta(P) = &\int_{A_0} f_\theta dP(\theta) + \sum_{j=1}^{K_0} \left[p_j - p_j^0\right] f_{\theta_j^0} + \int_{B(\theta_j^0, \epsilon)} \left(\theta - \theta_j^0\right)^T \nabla f_{\theta_j^0} dP(\theta) \\ &+ \frac{1}{2} \int_{B(\theta_j^0, \epsilon)} \left(\theta - \theta_j^0\right)^T D^2 f_{\theta_j^0} \left(\theta - \theta_j^0\right) dP(\theta)\end{aligned}$$

where for all $j = 1, \ldots, K_0$,

$$p_j = \sum_{i: \theta_i \in B(\theta_j^0, \epsilon)} \varpi_i$$

for some $\epsilon > 0$ and $\nabla f_{\theta_j^0}$ and $D^2 f_{\theta_j^0}$ denote respectively the gradient and the Hessian of f_θ evaluated at θ_j^0.
*Then, under Assumption **A2**, there exists a constant $c(f_0)$ depending only on f_0 such that*

$$\begin{aligned}\|\Delta(P)\|_1 \geq c(f_0) \Bigg[P(A_0) + \sum_{j=1}^{K_0} \Bigg\{ \left|p_j - \varpi_j^0\right| + \left\| \int_{B(\theta_j^0,\epsilon)} \left(\theta - \theta_j^0\right) dP(\theta) \right\| + \\ \int_{B(\theta_j^0,\epsilon)} \left\|\theta - \theta_j^0\right\|^2 dP(\theta) \Bigg\} \Bigg] \\ := c(f_0) N(P)\end{aligned} \tag{3.22}$$

where $B(\theta_j^0, \epsilon) = \{\theta : \|\theta - \theta_j^0\| \leq \epsilon\}$.

Proof. Let

$$\begin{aligned}\tilde{\Delta}(P) &= \left\| \frac{\Delta(P)}{N(P)} \right\|_1 \\ &= \left\| \alpha_0 \int_{A_0} f_\theta(x) d\nu_0(x) + \sum_{j=1}^{K_0} \alpha_j f_{\theta_j^0} + \beta_j^T \nabla f_{\theta_j^0} + \frac{1}{2} tr(D^2 f_{\theta_j^0} \gamma_j) \right\|_1\end{aligned}$$

where

$$\begin{cases} \alpha_0 & = \frac{P(A_0)}{N(P)}, \\ \nu_0(d\theta) & = \frac{P(d\theta)\mathbb{1}_{A_0}}{P(A_0)}, \\ \alpha_j & = \frac{p_j - p_j^0}{N(P)} \\ \beta_j & = \frac{1}{N(P)} \int_{B(\theta_j^0,\epsilon)} (\theta - \theta_j^0) dP(\theta), \\ \gamma_j & = \frac{1}{N(P)} \int_{B(\theta_j^0,\epsilon)} (\theta - \theta_j^0)(\theta - \theta_j^0)^T dP(\theta) \end{cases} \tag{3.23}$$

and assume that (3.22) does not hold. Then there exists a sequence $(P_m)_m$ with $\alpha_0^m, \nu_0^m, \alpha_j^m, \beta_j^m, \gamma_j^m$ for $j = 1, \ldots, K_0$ defined as in (3.23) along which $\tilde{\Delta}(P_m)$ goes to zero. Note that by construction

$$tr(\gamma_j) = \frac{1}{N(P)} \int_{B(\theta_j^0,\epsilon)} \|\theta - \theta_j^0\|^2 dP(\theta)$$

and that for $j = 1, \ldots, K_0$, $(\alpha_j^m, \beta_j^m, \gamma_j^m)_{m \in \mathbb{N}}$ belong to a compact set so that there exists a sub-sequence still labeled $(\alpha_j^m, \beta_j^m, \gamma_j^m)_{m \in \mathbb{N}}$ which is convergent to some value $(\alpha_j^*, \beta_j^*, \gamma_j^*)$.
Similarly, ν_0^m is a sequence of measure with mass bounded by 1, so it converges vaguely to a sub-probability measure ν_0^* on A_0 along a subsequence ν_0^m. Since for all $x \in \mathbb{R}^d$, $f_\theta(x)$ is continuous in θ and converges to 0 on the boundary of Θ,

$$\int_{A_0} f_\theta(x) d\nu_0^m(\theta) \overset{m \to \infty}{\longrightarrow} \int_{A_0} f_\theta(x) d\nu_0^*(\theta), \quad \forall x \in \mathbb{R}^d$$

Now

$$\int_{A_0} f_\theta(x) d\nu_0^m(\theta) \leq \sup_\theta \|f_\theta(\cdot)\|_\infty$$

hence on any compact subset B of $\mathbb{R}^d$,

$$\left\| \mathbb{1}_B(\cdot) \int_{A_0} f_\theta(x)(d\nu_0^m - d\nu_0^*)(\theta) \right\|_1 \overset{m\to\infty}{\longrightarrow} 0$$

so that for all compact subset B of $\mathbb{R}^d$ and for all $x \in B$,

$$\alpha_0^* \int_{A_0} f_\theta(x) d\nu_0^*(\theta) + \sum_{j=1}^{K_0} \alpha_j^* f_{\theta_j^0}(x) + \beta_j^{*T} \nabla f_{\theta_j^0}(x) + \frac{1}{2} tr(D^2 f_{\theta_j^0}(x) \gamma_j^*) = 0$$

for $j = 1, \ldots, K_0$.
Since the relation is true for all B, it is true for all $x \in \mathbb{R}^d$. The strong identifiability assumption **A2** implies that $\alpha_0^* = 0$, $\alpha_j^* = 0$, $\beta_j^* = 0$ and $\gamma_j^* = 0$, which is not possible since

$$\begin{aligned} \alpha_0^* + \sum_{j=1}^{K_0} |\alpha_j^*| + \|\beta_j^*\| + tr(\gamma_j^*) &= \lim_{m\to\infty} \alpha_0^m + \sum_{j=1}^{K_0} |\alpha_j^m| + \|\beta_j^m\| + tr(\gamma_j^m) \\ &= \lim_{m\to\infty} 1 = 1 \end{aligned}$$

Hence (3.22) is valid.

□

3.A.2 Proof of Theorem 3.1

To prove Theorem 3.1, we prove that the associated posterior concentrates in L_1 norm at the rate $\delta_n = (\log n)^q/\sqrt{n}$ for some $q > 0$ and then we bound from above $\Pi(\{P : \|f_P - f_0\|_1 \leq (\log n)^q/\sqrt{n}\})$. Let $D(K_0) = K_0 - 1 + dK_0$, if we denote by

$\ell_n(P) = \log f_p(\boldsymbol{y})$ the log likelihood, then for some $0 < t < \zeta$,

$$\begin{aligned}
&\mathbb{P}_{f_0}\left(\int e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\right)\\
&= \mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P)\right.\\
&\qquad \left.+\int_{\|f_P-f_0\|_1\le\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\right)\\
&\le \mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > \frac{1}{2}n^{-(D(K_0)+t)/2}\right)\\
&\qquad + \mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1\le\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > \frac{1}{2}n^{-(D(K_0)+t)/2}\right)
\end{aligned}$$

Using Markov's inequality and Fubini,

$$\begin{aligned}
&\le \mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > \frac{1}{2}n^{-(D(K_0)+t)/2}\right)\\
&\qquad + 2n^{(D(K_0)+t)/2}\Pi\left(P : \|f_P - f_0\|_1 \le \delta_n\right)
\end{aligned}$$

The most difficult part of the proof is to show that the right-most term converges to 0. It requires to control the a priori mass of decreasing neighborhoods of f_0, which is achieved in in Lemma 3.3. Therefore, it is enough to show that

$$\mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\right) = o(1). \tag{3.24}$$

Proof of (3.24)**.** Throughout the proof C denotes a generic constant depending only on $f_0 = f_{P_0}$. To prove (3.24), we use the strategy of the proof of Theorem 2.4 of Ghosal et al. 2000. Let $\eta > 0$ be arbitrarily small,

$$\mathcal{F}_n = \{f_P :\ P(\Theta_n^c) \le \eta\delta_n\}, \quad \Theta_n = \{\|\theta\| \le (\log n)^a\},$$

for some $a > 0$ and

$$\mathcal{F}_{n,\ell} = \{f_P \in \mathcal{F}_n; \|f_P - f_0\|_1 \in (\ell\delta_n, (\ell+1)\delta_n)\}, \quad \ell \ge 1.$$

Splitting

$$\begin{aligned}
\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) = \sum_{\ell\ge1}\int_{\mathcal{F}_{n,\ell}} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P)\\
+\int_{\mathcal{F}_n^c} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P),
\end{aligned}$$

we bound

$$
\begin{aligned}
&\mathbb{P}_{f_0}\left(\int_{\|f_P-f_0\|_1>\delta_n} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\right)\\
&\quad= P_{f_0}\left(\sum_{\ell\geq 1}\int_{\mathcal{F}_{n,\ell}} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) + \int_{\mathcal{F}_n^c} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\right)\\
&\quad\leq P_{f_0}\left(\sum_{\ell\geq 1}\int_{\mathcal{F}_{n,\ell}} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}/2\right)\\
&\qquad+ P_{f_0}\left(\int_{\mathcal{F}_n^c} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}/2\right)\\
&\quad\leq \sum_{\ell\geq 1} P_{f_0}\left(\int_{\mathcal{F}_{n,\ell}} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}\Pi(\mathcal{F}_{n,\ell})/2\right)\\
&\qquad+ P_{f_0}\left(\int_{\mathcal{F}_n^c} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}/2\right)\\
&\quad\leq \sum_{\ell\geq 1}\mathbb{E}_{f_0}\left\{\mathbb{1}\left(\int_{\mathcal{F}_{n,\ell}} f_P/f_0 d\Pi(P) > n^{-(D(K_0)+t)/2}\Pi(\mathcal{F}_{n,\ell})/2\right)(\phi_{n,\ell}+(1-\phi_{n,\ell}))\right\}\\
&\qquad+ P_{f_0}\left(\int_{\mathcal{F}_n^c} e^{\ell_n(P)-\ell_n(P_0)}d\Pi(P) > n^{-(D(K_0)+t)/2}/2\right)\\
&\quad\leq \sum_{\ell\geq 1}\left[\mathbb{E}_{f_0}(\phi_{n,\ell}) + \frac{2\int_{\mathcal{F}_{n,\ell}}\mathbb{E}_{f_P}(1-\phi_{n,\ell})d\Pi(P)}{\Pi(\mathcal{F}_{n,\ell})n^{-(D(K_0)+t)/2}}\right] + \frac{2\Pi(\mathcal{F}_n^c)}{n^{-(D(K_0)+t)/2}}
\end{aligned}
$$

where the last inequality uses Markov inequality and Fubini and where the $\phi_{n,\ell}$ are the L_1 tests in slice $\mathcal{F}_{n,\ell}$ as in Theorem 7.1 of Ghosal et al. 2000. In other words, if $1/18 \geq \omega > 0$, and if $B_{i,\ell}$, for $i \leq \mathcal{N}_{n,\ell}$, is a covering of $\mathcal{F}_{n,\ell}$ by L_1 balls of radius $\omega\ell\delta_n$, then $\phi_{n,i,\ell}$, $i \leq \mathcal{N}_{n,\ell}$ are the individual L_1 tests satisfying

$$\mathbb{E}_{f_0}(\phi_{n,i,\ell}) \leq e^{-c_1 n\ell^2\delta_n^2}, \quad \sup_{f\in B_{i,\ell}}\mathbb{E}_{f_P}(1-\phi_{n,i,\ell}) \leq e^{-c_1 n\ell^2\delta_n^2}$$

then $\phi_{n,\ell} = \max_{i=1}^{\mathcal{N}_{n,\ell}}\phi_{n,i,\ell}$ satisfies

$$\mathbb{E}_{f_0}(\phi_{n,\ell}) \leq \mathcal{N}_{n,\ell}e^{-c_1 n\ell^2\delta_n^2}, \quad \sup_{f_P\in\mathcal{F}_{n,\ell}}\mathbb{E}_{f_P}(1-\phi_{n,\ell}) \leq e^{-c_1 n\ell^2\delta_n^2}$$

Therefore to prove (3.24), it is enough to bound for some $\rho > 0$,

$$\mathcal{N}_{n,\ell} \lesssim e^{(\log n)^{2q-\rho}}, \tag{3.25}$$

and showing that

$$\Pi(\mathcal{F}_n^c) = o(n^{-(D(K_0)+t)/2}). \tag{3.26}$$

Using Markov inequality and assumption **A4**,

$$\begin{aligned}\Pi(P(\Theta_n^c) > \eta\delta_n) &\leq \frac{G_0(\|\theta\| \geq (\log n)^a)}{\eta\delta_n} \\ &= \frac{G_0(e^{a_1\|\theta\|^{a_2}} \geq e^{a_1(\log n)^{aa_2}})}{\eta\delta_n} \\ &\lesssim \frac{\exp(-a_1(\log n)^{aa_2})}{\delta_n} \\ &\lesssim n^{-(D(K_0)+2t)/2}\end{aligned}$$

by choosing $a > 1/a_2$ large enough so that the last equation holds, which in turn proves that (3.26) is verified. We now prove (3.25).

Using Lemma 3.7, we have that for all $f_P \in \mathcal{F}_{n,\ell}$, if $\ell\delta_n \leq \epsilon_0$ for some $\epsilon_0 > 0$ small enough,

$$P(A_0) + \sum_{j=1}^{K_0}\left[|p_j - p_j^0| + \|\mu(P_j) - \theta_j^0\| + \int_{B(\theta_j^0,\epsilon)} \|\theta - \theta_j^0\|^2 dP(\theta)\right] \leq \frac{(\ell+1)\delta_n}{c(f_0)} \tag{3.27}$$

where $\mu(P_j) := \int_{B(\theta_j^0,\epsilon)} \theta dP(\theta)/p_j$, $p_j = P(B(\theta_j^0, \epsilon))$.
Let $\xi > 0$, we consider a covering of $A_{0,n} = A_0 \cap \Theta_n$ with balls $U_{i,0}$ with center ϑ_i and radius ξ and $N_0(\xi)$, the number of such balls, is bounded by $(C|\Theta_n|/\xi)^d$. Define $\tilde{U}_{1,0} = U_{1,0}$ and $\tilde{U}_{i,0} = U_{i,0} \setminus \cup_{j=1}^{i-1} U_{j,0}$ for $i \geq 2$. Consider P, P' such that $f_P, f_P' \in \mathcal{F}_{n,\ell}$,

$$\begin{aligned}&\int \left|\int_{A_0} f_\theta(x)(dP - dP')(\theta)\right| dx \\ &\quad\leq \int \left|\sum_{i=1}^{N_0(\xi)} \left\{\int_{\tilde{U}_{i,0}} (f_\theta(x) - f_{\vartheta_i}(x))(dP - dP')(\theta)\right.\right. \\ &\qquad\left.\left. + \int_{\tilde{U}_{i,0}} f_{\vartheta_i}(x)(dP - dP')(\theta)\right\}\right| dx \\ &\quad\leq \int \sum_{i=1}^{N_0(\xi)} \left\{\int_{\tilde{U}_{i,0}} |f_\theta(x) - f_{\vartheta_i}(x)| d[P + P'](\theta)\right. \\ &\qquad\left. + f_{\vartheta_i}(x)\left|\int_{\tilde{U}_{i,0}} d(P - P')(\theta)\right|\right\} dx \\ &\quad\leq ||H_1(\cdot)||_1 \xi\,[P(A_0) + P'(A_0)] + \sum_{i=1}^{N_0(\xi)} \left|P(\tilde{U}_{i,0}) - P'(\tilde{U}_{i,0})\right| \\ &\quad\leq C\xi(\ell+1)\delta_n + \sum_{i=1}^{N_0(\xi)} \left|P(\tilde{U}_{i,0}) - P'(\tilde{U}_{i,0})\right|,\end{aligned}$$

where the third and last inequalities come from Assumption **A1** and equation (3.27) respectively.

Hence if for all $i \leq N_0(\xi)$, $|P(\tilde{U}_{i,0}) - P'(\tilde{U}_{i,0})| \leq \delta_n \xi / N_0(\xi)$,

$$\left\| \int_{A_0} f_\theta(\cdot)(dP - dP')(\theta) \right\|_1 \leq C\xi\ell\delta_n.$$

Moreover for each $j = 1, \ldots, K_0$, writing $P_j = P\mathbb{1}_{B(\theta_j^0,\epsilon)}/P(B(\theta_j^0,\epsilon))$ and $p_j = \sum_{i:\theta_i \in B(\theta_j^0,\epsilon)} \varpi_i$ and defining P'_j and p'_j in a similar way,

$$\begin{aligned}
\int_{B(\theta_j^0,\epsilon)} f_\theta(x)(dP - dP')(\theta) = \int_{B(\theta_j^0,\epsilon)} \Big\{ & f_{\theta_j^0}(x) + (\theta - \theta_j^0)^T \nabla f_{\theta_j^0}(x) \\
& + \frac{1}{2}(\theta - \theta_j^0)^T D^2 f_{\theta_j^0}(x)(\theta - \theta_j^0) \\
& + \mathcal{O}\left(\|\theta - \theta_j^0\|^{2+\delta}\right) \Big\} \left[dP - dP'\right](\theta) \\
= & \left[f_{\theta_j^0}(x) - \theta_j^{0T} \nabla f_{\theta_j^0}(x)\right] \left(p_j - p'_j\right) \\
& + \left[p_j \mu(P_j) - p'_j \mu(P'_j)\right]^T \nabla f_{\theta_j^0}(x) \\
& + \frac{1}{2} tr\left[D^2 f_{\theta_j^0}(x) \int (\theta - \theta_j^0)(\theta - \theta_j^0)^T d(P_j - P'_j)(\theta)\right] \\
& + \mathcal{O}\left(\int \|\theta - \theta_j^0\|^{2+\delta} d(P_j - P'_j)(\theta)\right).
\end{aligned}$$

Hence if $|p_j - p'_j| \leq \xi(\ell+1)\delta_n$, $\|\mu(P_j) - \mu(P'_j)\| \leq \xi(\ell+1)\delta_n$ and

$$\left\| \int (\theta - \theta_j^0)(\theta - \theta_j^0)^T d(P_j - P'_j)(\theta) \right\|_F \leq \xi(\ell+1)\delta_n,$$

$$\begin{aligned}
& \left\| \int_{B(\theta_j^0,\epsilon)} f_\theta(\cdot)(dP - dP')(\theta) \right\|_1 \\
& \quad \leq \xi(\ell+1)\delta_n \left(1 + \|H_1(\cdot)\|_1(\epsilon+1) + \frac{\|H_2(\cdot)\|_1}{2}\right) + \mathcal{O}(\epsilon^\delta \xi(\ell+1)\delta_n) \\
& \quad \leq C\xi\ell\delta_n.
\end{aligned}$$

So that $\|f_P - f_{P'}\|_1 \leq C\xi\ell\delta_n$.

Since for all j, when $f_P \in \mathcal{F}_{n,\ell}$, $\mu(P_j) \in B(\theta_j^0, C\ell\delta_n)$, the covering number for $\{\mu(P_j);\ f_P \in \mathcal{F}_{n,\ell}\}$ by balls of radius is $\xi\ell\delta_n$ is bounded by $C(\xi)^{-d}$ and similarly the covering numbers for the terms $\int(\theta - \theta_j^0)(\theta - \theta_j^0)^T dP_j(\theta)$ and p_j are bounded respectively by $C(\xi)^{-d(d+1)/2}$ and $C(\xi)^{-1}$.

This implies that by choosing ξ and small enough,

$$\begin{aligned}\mathcal{N}_{n,\ell} &\lesssim \xi^{-K_0[1+d+d(d+1)/2]}(N_0(\xi)/\xi)^{N_0(\xi)} \\ &\leq \exp\left[C\left(|\Theta_n|/\xi\right)^d \log(|\Theta_n|/\xi) + \log(1/\xi)\right] \\ &\leq \exp\left[C\left(|\Theta_n|/\xi\right)^d \left(\log(|\Theta_n|) + \log(1/\xi)\right)\right] \\ &\leq \exp\left[C\left((\log n)^a/\xi\right)^d \left(\log((\log n)^a) + \log(1/\xi)\right)\right]\end{aligned}$$

since $|\Theta_n| \leq (\log n)^a$ and (3.25) holds, by choosing q large enough such that $ad < 2q$, which in turns proves that (3.24) holds.

3.A.3 Proof of Corollary 3.2

By Theorem 3.1, if $f_0 = f_{P_0} \in \mathfrak{M}_{K_0}$, then for all $0 < t < \zeta$,

$$m_{DP}(\boldsymbol{y}) \lesssim n^{-D(K_0)/2 - t/2}$$

with P_{f_0}-probability going to 1. Since $\mathfrak{M}_{K_0}$ is regular we also have with probability going to 1,

$$m_{K_0}(\boldsymbol{y}) \gtrsim n^{-D(K_0)/2}$$

so that with probability going to 1,

$$\frac{m_{DP}(\boldsymbol{y})}{m_{K_0}(\boldsymbol{y})} \lesssim n^{-t/2} \underset{n\mapsto\infty}{\longrightarrow} 0 \quad \text{under } P_{f_0}.$$

Moreover using Rousseau and Mengersen 2011, for $K > K_0$, if $\alpha < d/2$, then

$$m_K(\boldsymbol{y}) \gtrsim n^{-D(K_0)/2 - (K-K_0)\alpha/2}$$

so that with probability going to 1, if $(K - K_0)\alpha < \zeta$,

$$\begin{aligned}\frac{m_{DP}(\boldsymbol{y})}{m_K(\boldsymbol{y})} &\lesssim n^{-t/2+(K-K_0)\alpha/2} \\ &\underset{n\mapsto\infty}{\longrightarrow} 0 \quad \text{under } P_{f_0} \text{ for } t \text{ such that } (K-K_0)\alpha < t < \zeta.\end{aligned}$$

On the other hand, for $K > K_0$ and $\alpha > d/2$, then

$$m_K(\boldsymbol{y}) \gtrsim n^{-D(K_0)/2 - d(K-K_0)/4}$$

so that with probability going to 1, if $d(K - K_0)/2 < \zeta$

$$\begin{aligned}\frac{m_{DP}(\boldsymbol{y})}{m_K(\boldsymbol{y})} &\lesssim n^{-t/2+d(K-K_0)/4} \\ &\underset{n\mapsto\infty}{\longrightarrow} 0 \quad \text{under } P_{f_0} \text{ for } t \text{ such that } d(K-K_0)/2 < t < \zeta.\end{aligned}$$

If $f_0 \notin \mathfrak{M}_K$ but $\Pi(KL(f_0, f_P) \leq \epsilon) > 0$ for all ϵ, then Ghosal et al. 2000 implies that for all $\epsilon > 0$, with probability going to 1

$$m_{DP}(\boldsymbol{y}) \gtrsim e^{-n\epsilon}$$

Since $\inf_{f_P \in \mathfrak{M}_K} KL(f_0, f_P) := c_K > 0$ then with probability going to 1

$$m_K(\boldsymbol{y}) \lesssim e^{-c_K n/2}$$

and Corollary 3.2 is proved.

3.A.4 Tables

	Tuning parameters		
	$n = 6$	$n = 36$	$n = 82$
Chib	$T_1 = 3 \cdot 10^4$	$T_1 = 5 \cdot 10^4$	$T_1 = 10^5$
	$burnIn = 2 \cdot 10^3$	$burnIn = 5 \cdot 10^3$	$burnIn = 10^4$
	$T_2 = 2 \cdot 10^3$	$T_2 = 2 \cdot 10^3$	$T_2 = 2 \cdot 10^3$
RLR-SIS	$T_1 = 3 \cdot 10^4$	$T_1 = 5 \cdot 10^4$	$T_1 = 10^5$
	$burnIn = 2 \cdot 10^3$	$burnIn = 5 \cdot 10^3$	$burnIn = 10^4$
	$T_2 = 2 \cdot 10^3$	$T_2 = 2 \cdot 10^3$	$T_2 = 2 \cdot 10^3$
RLR-Prior	$T_1 = 3 \cdot 10^4$	$T_1 = 5 \cdot 10^4$	$T_1 = 10^5$
	$burnIn = 2 \cdot 10^3$	$burnIn = 5 \cdot 10^3$	$burnIn = 10^4$
	$T_2 = 2.8 \cdot 10^4$	$T_2 = 4.5 \cdot 10^4$	$T_2 = 9 \cdot 10^4$
Arithmetic Mean	$T = 3 \cdot 10^4$	$T = 2 \cdot 10^5$	$T = 5 \cdot 10^5$

Table 3.1: Choice of tuning parameters for Figure 3.3

Chapter 4

Distributed and in parallel evidence computation

Abstract

In this chapter, we consider the challenge of distributing the computation of the marginal likelihood of a finite mixture model. Buchholz et al. 2022 provides a very convenient solution to the problem of computing the marginal likelihood of a model when batches of data are distributed across several computing servers. Unfortunately, their approach cannot be applied to finite mixtures and we here develop *ad hoc* solutions for using their method on such models, as well as a new approach based on Sequential Monte Carlo. The latter is very promising given it does not rely on any conjugacy assumption and can be applied to other models that finite mixtures. We also provide an empirical study that highlights the decrease in computational time that is gained with respect to classical marginal likelihood estimates computed on the full dataset.

Contents

4.1 Introduction

Distributed computation has emerged as a powerful paradigm for tackling complex problems in machine learning and statistical inference, where the amount of data is too large to be processed by a single machine. This approach involves breaking down the inference task into smaller sub-problems that can be solved independently by multiple machines or processors, and then combining the results to obtain a global solution to the initial problem. This not only reduces the computational burden on individual machines but also facilitates parallel computation, enabling faster and more efficient inference.

From a Bayesian inference perspective, this strategy, usually called *divide and conquer*, is mainly composed of three steps. First, the data $\boldsymbol{y}$ is *divided* into S non-overlapping batches $\boldsymbol{y}_1, \dots, \boldsymbol{y}_S$ and the full posterior distribution is decomposed as

$$\begin{aligned} \pi(\boldsymbol{\vartheta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) \\ &= \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} \\ &\propto \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) \end{aligned} \tag{4.1}$$

for a likelihood function $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ and a prior $\pi(\boldsymbol{\vartheta})$, where for all $s = 1, \dots, S$, $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ is called the sub-posterior distribution of parameter $\boldsymbol{\vartheta}$ on batch $\boldsymbol{y}_s$. Posterior inference is then conducted on each subset of the data independently, possibly in parallel, running MCMC algorithms across several computing units. Lastly, the different sub-posterior samples are *recombined* to approximate a sample from the *full* posterior distribution. This last step is usually complex as there exists no exact way to transform a collection of samples $\{\boldsymbol{\vartheta}_t^{(s)}\}_t$ from each sub posterior $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ into a sample $\{\boldsymbol{\vartheta}_t\}_t$ that is distributed according to $\pi(\boldsymbol{\vartheta}|\boldsymbol{y})$. Based on the Bernstein-von Mises theorem, Huang and Gelman 2005 and Scott et al. 2016 use normal approximations to the sub-posterior distributions to reconstruct a sample from the full posterior as a weighted average of the sub-posterior samples. Still assuming normality of the sub-posterior distributions, Neiswanger et al. 2014 uses kernel density estimation to reconstruct the full posterior. Other approaches consist in recombining the sub-samples through their barycenter in a Wasserstein space of probability measures (Srivastava et al. 2018), or their geometric median (Minsker et al. 2014). Applications of a *divide and conquer* strategy are numerous. First and foremost, it can be used in order to significantly reduce the cost of posterior inference when the amount of data at hand is very large by reducing memory and computational bottle necks. Another immediate application is inference on a distributed data architecture, in which data is stored at different locations, either for privacy issues, for instance with health data (Hallock et al. 2021), or simply because the data set size is too large to be stored on a single machine.

Despite its popularity, the application of the *divide and conquer* paradigm to mixture modeling remains largely unexplored so far. Indeed, while conducting

inference on each data batch $\boldsymbol{y}_s$ is generally not an issue, the usual assumption of asymptotic normality used for recombination of the S MCMC samples does not hold for the posterior distribution of mixture models.

From a Bayeseian model selection perspective, the issue of computing the marginal likelihood of the data $\boldsymbol{y}$, defined as

$$m(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\vartheta})ß(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}$$

is crucial. Indeed, the Bayesian paradigm offers a straightforward way of comparing two competing models $\mathfrak{M}_0$ and $\mathfrak{M}_1$ through the Bayes Factor (Jeffreys 1935), defined as the ratio of the marginal likelihoods of the data under each model. The issue of distributed computation of the marginal likelihood of a model in a *divide and conquer* fashion remains largely unexplored so far. The issue being that while there exists a convenient way of linking the full posterior distribution and the sub-posterior distribution through (4.1), such an identity does not hold for the marginal likelihood. Indeed,

$$\begin{aligned} m(\boldsymbol{y}) &= \int \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta} \\ &\neq \prod_{s=1}^{S} \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}. \end{aligned}$$

This calls for a different approach to combining marginal likelihoods estimates computed on each batch $\boldsymbol{y}_s$ separately. Very recently, an identity bridging the gap between $m(\boldsymbol{y})$ and $\{\tilde{m}(\boldsymbol{y}_s)\}_{s=1}^{S}$ has been derived by Buchholz et al. 2022, where $\tilde{m}(\boldsymbol{y}_s)$ is the marginal likelihood associated to posterior $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$. Unfortunately, this identity is not suited for the intrinsic complexity of finite mixture models and in particular their lack of identifiability. In this chapter, we first propose an adaptation of the framework set by Buchholz et al. 2022 before deriving a new identity linking the full data marginal likelihood and the batch marginal likelihoods. This new identity enables the implementation of a Sequential Monte Carlo strategy for a fully distributed computation of the marginal likelihood.

4.2 A simple identity for distributed computation of marginal likelihoods

4.2.1 Notation

Assume that some data $\boldsymbol{y} \in \mathbb{R}^d$ can be splitted into S non-overlapping, non-empty batches $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)$. For a finite mixture model with K components as specified in Chapter 2, given the component parameters $\boldsymbol{\theta} = (\theta, \ldots, \theta_K) \in \mathbb{R}^{dK}$ and weights $\boldsymbol{\varpi} = (\varpi_1, \ldots, \varpi_K) \in \Delta_{K-1}$, with Δ_{K-1} the $(K-1)$-dimensional simplex, the likelihood $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ can be factorized as

$$p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\varpi})$. The posterior density can thus be derived up to a constant as

$$\begin{aligned}\pi(\boldsymbol{\vartheta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) \\ &= \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S} \\ &\propto \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|y_s)\end{aligned}$$

where for all $s = 1, \ldots, S$, the densities $\tilde{\pi}(y_s|\boldsymbol{\vartheta})$ are the so-called *sub-posterior* densities, that is the posterior distribution arising from the batched likelihood $p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$ and the prior distribution $\tilde{\pi}(\boldsymbol{\vartheta}) \propto \pi(\boldsymbol{\vartheta})^{1/S}$. The following section presents the work of Buchholz et al. 2022 which provides a useful expression of the marginal likelihood on the full data $m(\boldsymbol{y})$ as a function of the batched marginal likelihoods $\{\tilde{m}(\boldsymbol{y}_s)\}_{s=1}^{S}$.

4.2.2 An identity by Buchholz et al. 2022

Proposition 4.1 below is derived in Buchholz et al. 2022 and provides a convenient bridge between the batched marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ and the full marginal likelihood $m(\boldsymbol{y})$.

Proposition 4.1 (Buchholz et al. 2022)**.** *For some data $\boldsymbol{y}$ and a model $\mathcal{P}$ for which the likelihood function factorizes as $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$, the marginal likelihood of the data can be written as*

$$m(\boldsymbol{y}) = Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4.2}$$

where for each $s = 1, \ldots, S$,

$$\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) \propto p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta}),$$
$$\tilde{m}(\boldsymbol{y}_s) = \int p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta})d\boldsymbol{\vartheta},$$

and

$$Z = \int \pi(\boldsymbol{\vartheta})^{1/S} d\boldsymbol{\vartheta}$$

Proof. Using Bayes formula,

$$\begin{aligned}
\pi(\boldsymbol{\vartheta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{m(\boldsymbol{y})} \\
&= \frac{\prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})^{1/S}}{m(\boldsymbol{y})} \\
&= Z^S \frac{\prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})\tilde{\pi}(\boldsymbol{\vartheta})}{m(\boldsymbol{y})} \\
&= Z^S \frac{\prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)}{m(\boldsymbol{y})}
\end{aligned}$$

Integrating both sides with respect to $\boldsymbol{\vartheta}$ and rearranging,

$$m(\boldsymbol{y}) = Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$$

□

Roughly speaking, identity (4.2) expresses the marginal likelihood on the full data as the product of the batched marginal likelihoods, corrected by a measure of similarity of the batches $\{\boldsymbol{y}_s\}$, measured by the normalizing constant of the product of the sub-posterior distributions.
Computing the prior normalizing constant Z is typically straightforward for the most common choices of mixture priors, provided the hyper parameters are chosen carefully. Example below gives Z when the mixture weights follow the Dirichlet distribution $\mathcal{D}(\cdot|\boldsymbol{\alpha})$ a priori and when the mixture components are assumed to arise from a Normal-Inverse-Gamma $\mathcal{NIG}(\cdot|\mu_0, \lambda, a, b)$.

Proposition 4.2. *Consider a* $K-$*component mixture model with prior distribution on the component weights and parameters* $(\boldsymbol{\varpi}, \boldsymbol{\theta})$ *where* $\boldsymbol{\theta} = ((\mu_1, \sigma_1^2), \ldots, (\mu_K, \sigma_K^2))$ *given by*

$$\Pi(d\boldsymbol{\varpi}, d\boldsymbol{\theta}) = \mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}) \times \prod_{k=1}^{K} \mathcal{NIG}(d\boldsymbol{\theta}_k|\mu_0, \lambda, a, b)$$

with density π. *Then, for all integer* $S \geq 1$, $\tilde{\pi}(\boldsymbol{\vartheta}) \propto \pi(\boldsymbol{\vartheta})^{1/S}$ *is distributed as*

$\mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}') \times \prod_{k=1}^{K} \mathcal{NIG}(d\boldsymbol{\theta}_k|\mu_0', \lambda', a', b')$ *where*

$$\begin{cases} \alpha_k' & = (\alpha_k - 1 + S)/S \ \ \textit{for } k = 1, \ldots, K \\ \mu_0' & = \mu_0 \\ \lambda' & = \lambda/S \\ a' & = (a + 3/2 - (3/2)S)/S \\ b' & = b/S \end{cases}$$

provided $a > (3/2)(S-1)$.

Proof. Proof in Appendix. □

Corollary 4.3. *For a* $K-$*component mixture of Normal distributions with prior distribution on the component weights and parameters* $(\boldsymbol{\varpi}, \boldsymbol{\theta})$ *where* $\boldsymbol{\theta} = ((\mu_1, \sigma_1^2), \ldots, (\mu_K, \sigma_K^2))$ *given by*

$$\Pi(d\boldsymbol{\varpi}, d\boldsymbol{\theta}) = \mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}) \times \prod_{k=1}^{K} \mathcal{NIG}(d\theta_k|\mu_0, \lambda, a, b)$$

with density π*, the normalizing constant* Z *of* $\pi^{1/S}$ *for some integer* $S \geq 1$*, is given by*

$$\begin{aligned} Z = & \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^{1/S} \frac{\prod_{k=1}^{K} \Gamma((\alpha_k - 1 + S)/S}{\Gamma((1/S) \sum_{k=1}^{K} \alpha_k - 1 + S)} \\ & \times \left(\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \right)^{1/S} \frac{\sqrt{2\pi}}{\sqrt{\lambda/S}} \frac{\Gamma((a + 3/2 - (3/2)S)/S)}{(b/S)^{(a+3/2-(3/2)S)/S}} \right)^{K} \end{aligned}$$

provided $a > (3/2)(S-1)$.

Proof. Noticing that the normalizing constant of $\pi^{1/S}$ is given by

$$\begin{aligned} \int \pi(\boldsymbol{\varpi}, \boldsymbol{\theta})^{1/S} d\boldsymbol{\varpi} d\boldsymbol{\theta} &= \left(\int \pi(\boldsymbol{\varpi})^{1/S} d\boldsymbol{\varpi} \right) \times \left(\int \pi(\boldsymbol{\theta})^{1/S} d\boldsymbol{\theta} \right) \\ &= \left(\int \pi(\boldsymbol{\varpi})^{1/S} d\boldsymbol{\varpi} \right) \times \prod_{k=1}^{K} \left(\int \pi(\boldsymbol{\theta}_k)^{1/S} d\boldsymbol{\theta}_k \right) \end{aligned}$$

the result is an immediate consequence of Proposition 4.2 above □

Corollary 4.4. *For a* $K-$*component mixture of Normal distribution with prior distribution on the component weights and parameters* $(\boldsymbol{\varpi}, \boldsymbol{\theta})$ *where* $\boldsymbol{\theta} = ((\mu_1, \sigma_1^2), \ldots, (\mu_K, \sigma_K^2))$ *given by*

$$\Pi(d\boldsymbol{\varpi}, d\boldsymbol{\theta}) = \mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}) \times \prod_{k=1}^{K} \mathcal{NIG}(d\theta_k|\mu_0, \lambda, a, b)$$

then for all $s = 1, \dots, S$*, the augmented sub-posterior distribution of* $\boldsymbol{\vartheta}$ *is*

$$\tilde{\Pi}(d\boldsymbol{\varpi}, d\boldsymbol{\theta}|\boldsymbol{z}_s, \boldsymbol{y}_s) = \mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}_s) \times \prod_{k=1}^{K} \mathcal{NIG}(d\theta_k|\mu_{0sk}, \lambda_{sk}, a_{sk}, b_{sk})$$

where

$$\begin{cases} \alpha_{sk} = \alpha'_k + N_k(\boldsymbol{z}_s) \\ \mu_{0sk} = (N_k(\boldsymbol{z}_s)\overline{\boldsymbol{y}_{sk}} + \lambda'\mu'_0)/(N_k(\boldsymbol{z}_s) + \lambda') \\ \lambda_{sk} = \lambda' + N_k(\boldsymbol{z}_s) \\ a_{sk} = a' + N_k(\boldsymbol{z}_s)/2 \\ b_{sk} = b' + (1/2)\left[\sum_{i:z_{si}=k}(y_{si} - \overline{\boldsymbol{y}_{sk}})^2 + N_k(\boldsymbol{z}_s)\lambda'/(N_k(\boldsymbol{z}_s) + \lambda')(\overline{\boldsymbol{y}_{sk}} - \mu'_0)^2\right] \end{cases}$$

where $(\boldsymbol{\alpha}', \mu'_0, \lambda', a', b')$ *are the hyper-parameters of* $\tilde{\pi}(\boldsymbol{\vartheta})$ *derived in Proposition 4.2,* $N_k(\boldsymbol{z}_s) = |\{i : z_{si} = k\}|$ *and* $\boldsymbol{y}_{sk} = \{y_{si} : z_{si} = k\}$.

Proof. This result is an immediate consequence of the conjugacy of the prior derived in Proposition 4.2 and of Example 1 in Chapter 2, in which the posterior hyper-parameters are computed. □

Proposition 4.2 above shows that the shape parameter $a > 0$ of the popular Normal-Inverse Gamma prior is constrained to increase as the number of batches S increases.

The product of the marginal likelihoods on the batches of data in the identity (4.2) can usually be easily estimated. For instance, for a normal mixture kernel F and a Dirichlet-Normal-Inverse Gamma prior on the weights and parameters $\boldsymbol{\vartheta} = (\boldsymbol{\varpi}, \boldsymbol{\theta})$, Proposition 4.2 guarantees that estimating $\tilde{m}(\boldsymbol{y}_s)$ boils down to the problem of estimating the marginal likelihood of a finite mixture model with a conditionally conjugate prior for a suitable reparametrization, which was extensively studied in Chapter 2.

The right-most term of equation (4.2), however, is typically intractable for mixture models since the posterior density of the parameters and weights $\boldsymbol{\vartheta}$ cannot be written in closed form. Hence, the quantity

$$I = \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta}$$

is also intractable since it is the normalizing constant of a distribution proportional to the product of S mixture posteriors. To address this issue, Buchholz et al. 2022 gives an estimate of I that is available for general *conditionally conjugate* parametric models. As discussed in Chapter 2, finite mixtures of Normal with a Normal-Inverse Gamma prior belong to this class of models, for which the augmented posterior $\pi(\boldsymbol{\vartheta}|\boldsymbol{z}, \boldsymbol{y})$ where $\boldsymbol{z}$ is a latent variable (the cluster allocations in the case of mixtures), is available in closed-form. Using this convenient property of conditionally-conjugate models, Buchholz et al. 2022 suggests the following estimator of I given in Proposition 4.5 below.

Proposition 4.5 (Buchholz et al. 2022)**.** *For general conditionally conjugate models, the integral* $I := \int_\Theta \prod_{s=1}^S \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)d\boldsymbol{\vartheta}$ *can be estimated through*

$$\hat{I} = \frac{1}{T}\sum_{t=1}^{T}\int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s^{(t)}, \boldsymbol{y}_s)d\boldsymbol{\vartheta} \tag{4.3}$$

where $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ *and* $\hat{I}$ *is an unbiased estimator of* I*.*

Proof. Using the latent variable representation, we write

$$\prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) = \prod_{s=1}^{S} \int \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_s$$

By independence,

$$= \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_1 \dots d\boldsymbol{z}_S$$

Integrating both sides with respect to $\boldsymbol{\vartheta}$, we get

$$I = \int\int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_1 \dots d\boldsymbol{z}_S d\boldsymbol{\vartheta}$$

which yields by Fubini

$$\begin{aligned} I &= \int \left(\int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)d\boldsymbol{\vartheta}\right) \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_1 \dots d\boldsymbol{z}_S \\ &= \mathbb{E}_{\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})}\left[\int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)d\boldsymbol{\vartheta}\right] \end{aligned}$$

where we define $\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S}) := \prod_{s=1}^S \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$.
The proof is then concluded by a simple Monte-Carlo argument. □

This is summarized in Algorithm 14, which again holds for any *conditionally conjugate* parametric model. A full derivation of the augmented sub-posterior distribution $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s, \boldsymbol{y}_s)$ is given in Proposition 4.9 in Appendix.

Unfortunately, despite their conditionally conjugate nature, we highlight in the next section why the data augmentation trick for estimating I proposed by Buchholz et al. 2022 is not suited to conditionally conjugate finite mixture models.

4.2.3 Unapplicability to conditionally conjugate finite mixtures

To understand why the estimator $\hat{I}$ cannot be directly applied to conditionally conjugate finite mixtures, recall that, as extensively discussed in Chapter 2, the values taken by the latent cluster allocation variables z_i for $i = 1, \dots, n$ are completely non-informative when considered individually. That is, the numerical value taken

Algorithm 14 : Distributed marginal likelihood computation for conditionally conjugate parametric models (Buchholz et al. 2022)

Input : For all $s = 1, \dots, S$, $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_{t=1}^T$ samples from a MCMC sampler with stationary distribution $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}_s|\boldsymbol{y}_s)$, possibly run in parallel.

Do in parallel

  Estimate $\tilde{m}(\boldsymbol{y}_s)$, possibly using the sample $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_t$, for $s = 1, \dots, S$;

end

Compute $\hat{I}$ using (4.3);

Return $\hat{m}_{S,\hat{I}} = Z^S \prod_{s=1}^S \widehat{\tilde{m}(\boldsymbol{y}_s)} \times \hat{I}$;

by a single z_i is irrelevant if taken out of the context of the whole vector of allocations $\boldsymbol{z} = (z_1, \dots, z_n)$. In fact, for any permutation σ of the set $\{1, \dots, K\}$, $\tilde{\boldsymbol{z}} = (\sigma(z_1), \dots, \sigma(z_n))$ induces the same clustering structure of the observations as $\boldsymbol{z}$ and thus conveys exactly the same information. This leads to a *permutation invariance* of the posterior distribution of the allocations $\boldsymbol{z}$ which severely hinders the estimation of I through $\hat{I}$. Indeed, for all $t = 1, \dots, T$, where T is the number of simulations performed in each MCMC algorithm targeting one of the S sub-posterior distributions, estimator $\hat{I}$ requires to evaluate the integral

$$\int_\Theta \prod_{s=1}^S \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s^{(t)}, \boldsymbol{y}_s) d\boldsymbol{\vartheta}$$

where for all $s = 1, \dots, S$, the vector $\boldsymbol{z}_s^{(t)}$ is distributed according to $\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$.
While $\hat{I}$ remains an unbiased estimator of I for finite mixture models, its variance heavily depends on the properties of the MCMC chains targeting the mixture sub-posteriors $\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ and in particular the presence of a *label switching* phenomenon (or rather lack thereof), as illustrated on Figure 2.4 of Chapter 2 and its subsequent discussion. In an ideal scenario, the Gibbs sampler would visit evenly all $K!$ modal configurations of $\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$, since they are all equiprobable a posteriori for an exchangeable prior distribution. This is hardly ever the case for a finite number of simulations T. Moreover, this idealistic behavior of the Gibbs sampler should happen for all S MCMC chains, which is an even more demanding assumption.

To illustrate the inefficiency of $\hat{I}$ for finite mixtures we provide the following example.

Example 1. We generate $n = 2000$ *i.i.d* data points $\boldsymbol{y}$ from the following two-component mixture of Normal distributions $0.5\mathcal{N}(0.5, 1) + 0.5\mathcal{N}(3, 1)$ and use a location-scale finite normal mixture for $K = 1, 2, 3, 4$ and 5. For the full model (i.e using the full data $\boldsymbol{y}$), we choose the conditionally conjugate Dirichlet Normal Inverse-Gamma prior as specified in Proposition 4.2 with hyperparameters $\alpha_k = 1, \mu_{0k} = \bar{\boldsymbol{y}}, \lambda_k = 2.6/(\boldsymbol{y}_{(n)} - \boldsymbol{y}_{(1)}), a_k = 4, b_k = 0.36(\overline{\boldsymbol{y}^2} - \overline{\boldsymbol{y}}^2)$, for all $k = 1, \dots, K$. We choose to split the data into $S = 3$ batches and use the prior $\tilde{\pi}(\boldsymbol{\vartheta}) \propto \pi(\boldsymbol{\vartheta})^{1/3}$ which, by Proposition 4.2, is still the product of a Dirichlet distribution and a Normal Inverse-Gamma with a suitable reparametrization. The simulation results are presented in Figure 4.1. Note that each repetition used to create the boxplots

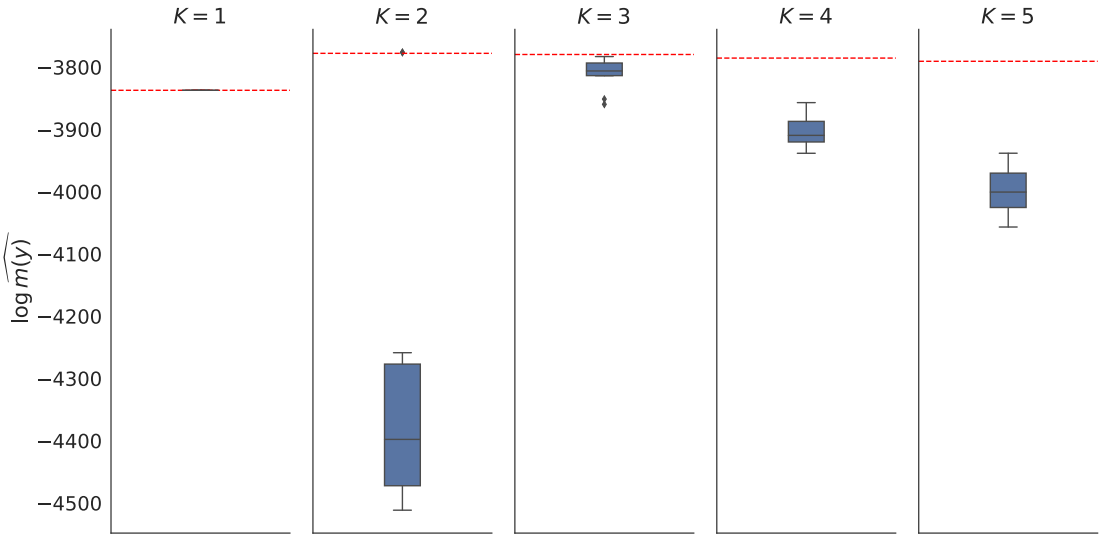


Figure 4.1: Example 1 : Boxplots of 10 repetitions of distributed marginal likelihood estimates for different values of K where I is estimated through $\hat{I}$ and the batch marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ are estimated by the permuted Chib's estimator $\hat{m}_{ChibPerm}(\boldsymbol{y}_s)$ (cf Equation (2.22) of Chapter 2). For each K, the reference red dashed line is computed on the whole dataset $\boldsymbol{y}$ using $\hat{m}_{ChibPerm}(\boldsymbol{y})$.

are done with a different random partition of the data $\boldsymbol{y}$ into 3 batches $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)$, so that the observed variance is due both to the Monte Carlo sampler and the effect of using different partitions of the full data $\boldsymbol{y}$.

For $K = 1$, we see that the distributed marginal likelihood estimator using $\hat{I}$ is perfectly aligned with the reference line, which is expected since the latent allocation variables $\boldsymbol{z}$ can only take one value and hence the labels of the $\boldsymbol{z}_s$ are necessarily coherent across the batches. However, as soon as $K > 1$, a pathological variance can be observed, leading to a downward bias in the log-scale.

It is interesting to remark that the variance is the greatest for the case $K = 2$, which in fact is the chosen number of components in the Data Generating Process (DGP). At first, this can seem quite paradoxical since well-specified mixtures are usually more easily estimated. However, in this situation, while it is true that within each batch the Gibbs sampler shows good mixing properties, a lack of label switching implies that sub-posterior samples typically represent a single modal configuration. If this modal configuration is coherent across the 3 batches, then $\hat{I}$ is a good approximation of I. The outlier point that lies almost perfectly on the reference line on the plot $K = 2$ is an example of this phenomenon, which happens in about $(1/K!)^S = (1/2)^3$ of the scenarios for this well-specified finite mixture example with rather distant components. However, the remaining cases in which the Gibbs samplers are not label-consistent across batches pay a heavy penalization through the integral $\int_\Theta \prod_{s=1}^S \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_s^{(t)}, \boldsymbol{y}_s)$, which explains the low value of the estimate of the marginal likelihood for $K = 2$. The ill-specified cases $K = 3, 4$ and 5 are more prone to label switching and are less penalized by the integral above for incoherent labels, which explains their 'better' performance.

What Example 1 helps us understand is that a perfect mixing of the MCMC

algorithms targeting the sub-posterior distributions is necessary for $\hat{I}$ to be a valuable estimate of I. By perfect mixing, it is implied that a balanced label-switching phenomenon should take place both within and across each S MCMC chains which is an unrealistic expectation. Nevertheless, this intuition is at the core of the idea behind the estimator $\hat{I}_{perm}$ that we derive in the next section, that enforces a perfect label switching phenomenon to improve the performance of $\hat{I}$.

4.3 Permuted estimator of I

Keeping in mind the previous considerations, and remembering the mixture posterior permutation-invariance explained in Section 2.2.3 of Chapter 2, an easy fix can be derived using the identity

$$\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) = \tilde{\pi}((\vartheta_{\sigma(1)}, \ldots, \vartheta_{\sigma(K)})|\boldsymbol{y}_s)$$

for all s and all permutation σ of the set $\{1, \ldots, K\}$. Hence, trivially,

$$\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) = \frac{1}{K!} \sum_{\sigma \in \mathfrak{S}_K} \tilde{\pi}((\vartheta_{\sigma(1)}, \ldots, \vartheta_{\sigma(K)})|\boldsymbol{y}_s) \tag{4.4}$$

where $\mathfrak{S}_k$ denotes the set of all permutations of $\{1, \ldots, K\}$. Equation (4.4) makes it possible to consider all possible modal configurations of the sub-posterior distributions. Without loss of generality, if we set batch 1 as reference and consider all possible permutations across all $S-1$ remaining batches, then the cluster label matching issue described above can be addressed. The trick here is not to consider separately all possible permutations $\sigma_s \in \mathfrak{S}_k$ of the labels of batch $s = 2, \ldots, S$, but rather to consider them jointly as a vector $\boldsymbol{\sigma} = (\sigma_2, \ldots, \sigma_S)$ on the product space $\mathfrak{S}_K^{S-1}$. Such a vector $\boldsymbol{\sigma}$ shall be called a *configuration* thereafter, or in other words, a combination of permutations.

Hence we introduce the following fully permuted estimator which effectively circumvents the issue of cluster labeling.

Proposition 4.6 (Permuted estimator of I)**.** *Let S be an integer larger than 1 and let $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ for $s = 1, \ldots, S$. Then*

$$\hat{I}_{perm} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2, \ldots, \sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_1^{(t)}, \boldsymbol{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s^{(t)}), \boldsymbol{y}_s) d\boldsymbol{\vartheta} \tag{4.5}$$

is an unbiased estimator for I.

Proof. Using the posterior permutation invariance,

$$\begin{aligned}\prod_{s=1}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) &= \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_1)\prod_{s=2}^{S}\frac{1}{K!}\sum_{\sigma_s\in\mathfrak{S}_K}\tilde{\pi}((\vartheta_{\sigma_s(1)},\ldots,\vartheta_{\sigma_s(K)})|\boldsymbol{y}_s)\\ &= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}((\vartheta_{\sigma_s(1)},\ldots,\vartheta_{\sigma_s(K)})|\boldsymbol{y}_s)\\ &= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\left(\int\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_1,\boldsymbol{y}_1)\tilde{\pi}(\boldsymbol{z}_1|\boldsymbol{y}_1)d\boldsymbol{z}_1\right)\\ &\quad\times\prod_{s=2}^{S}\left(\int\tilde{\pi}((\vartheta_{\sigma_s(1)},\ldots,\vartheta_{\sigma_s(K)})|\boldsymbol{z}_s,\boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_s\right)\end{aligned}$$

by independence and noticing that $\tilde{\pi}((\vartheta_{\sigma_s(1)},\ldots,\vartheta_{\sigma_s(K)})|\boldsymbol{z}_s,\boldsymbol{y}_s)=\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s),\boldsymbol{y}_s)$,

$$\begin{aligned}&= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\int\pi(\boldsymbol{\vartheta}|\boldsymbol{z}_1,\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s),\boldsymbol{y}_s)\\ &\quad\times\prod_{s=1}^{S}\pi(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_1,\ldots,d\boldsymbol{z}_s\end{aligned}$$

integrating both sides with respect to $\boldsymbol{\vartheta}$ and using Fubini's theorem,

$$\begin{aligned}I &= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\int\left(\int\pi(\boldsymbol{\vartheta}|\boldsymbol{z}_1,\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s),\boldsymbol{y}_s)d\boldsymbol{\vartheta}\right)\\ &\quad\times\prod_{s=1}^{S}\pi(\boldsymbol{z}_s|\boldsymbol{y}_s)d\boldsymbol{z}_1,\ldots,d\boldsymbol{z}_s\\ &= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\mathbb{E}_{\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})}\left[\int\pi(\boldsymbol{\vartheta}|\boldsymbol{z}_1,\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s),\boldsymbol{y}_s)d\boldsymbol{\vartheta}\right].\end{aligned}$$

Thus,

$$\begin{aligned}\hat{I}_{perm} &= \frac{1}{K!^{S-1}}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\frac{1}{T}\sum_{t=1}^{T}\int\pi(\boldsymbol{\vartheta}|\boldsymbol{z}_1^{(t)},\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s^{(t)}),\boldsymbol{y}_s)d\boldsymbol{\vartheta}\\ &= \frac{1}{TK!^{S-1}}\sum_{t=1}^{T}\sum_{\sigma_2,\ldots,\sigma_S\in\mathfrak{S}_K}\int\pi(\boldsymbol{\vartheta}|\boldsymbol{z}_1^{(t)},\boldsymbol{y}_1)\prod_{s=2}^{S}\tilde{\pi}(\boldsymbol{\vartheta}|\sigma_s(\boldsymbol{z}_s^{(t)}),\boldsymbol{y}_s)d\boldsymbol{\vartheta}\end{aligned}$$

where $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^{T}\overset{i.i.d}{\sim}\tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ for all $s=1,\ldots,S$ is an unbiased estimator for I. □

Algorithm 15 below gives a pseudo-code implementation of $\hat{I}_{perm}$.

Algorithm 15 : Distributed fully permuted marginal likelihood computation for conditionally conjugate finite mixture models

Input : For all $s = 1, \ldots, S$, $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_{t=1}^T$ samples from a MCMC sampler with stationary distribution $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}_s|\boldsymbol{y}_s)$, possibly run in parallel.

Do in parallel

  Estimate $\tilde{m}(\boldsymbol{y}_s)$, possibly using the sample $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_t$, for $s = 1, \ldots, S$;

end

Compute $\hat{I}_{perm}$ using (4.5);

Return $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y}) := Z^S \prod_{s=1}^S \widehat{\tilde{m}(\boldsymbol{y}_s)} \times \hat{I}_{perm}$;

Example 1. ***(continued)*** Reusing the same 2-component Normal mixture setting with $S = 3$ batches and its numerical simulations, we add estimates of the distributed marginal likelihood estimates using $\hat{I}_{perm}$. For the sake of easing comparison, we choose the same number of Gibbs sub-posterior simulations T for each method. More precisely, T is set to 1000, 51000 and 101000 for $K = 1, 2$ and 3 respectively, with a burn-in of 5%.

The improvement in the estimation of I through $\hat{I}_{perm}$ over $\hat{I}$ is tremendous, as shown on Figure 4.2. This is especially true for the well-specified case $K = 2$ in which the marginal likelihood estimates lie almost exactly on the reference line. This is also true for the case $K = 3$. However, this great performance come at a heavy computational cost, as expected. Figure 4.3 illustrates this phenomenon by plotting the log of the computational time required to get one of the repetitions of the boxplots for each K. Remember that the cost of estimating $\hat{I}$ is $\mathcal{O}(T)$ while it is $\mathcal{O}(TK!^{S-1})$ for $\hat{I}_{perm}$. This explains why the computational times are similar for $K = 1$, but differ greatly for $K = 2$ and 3.

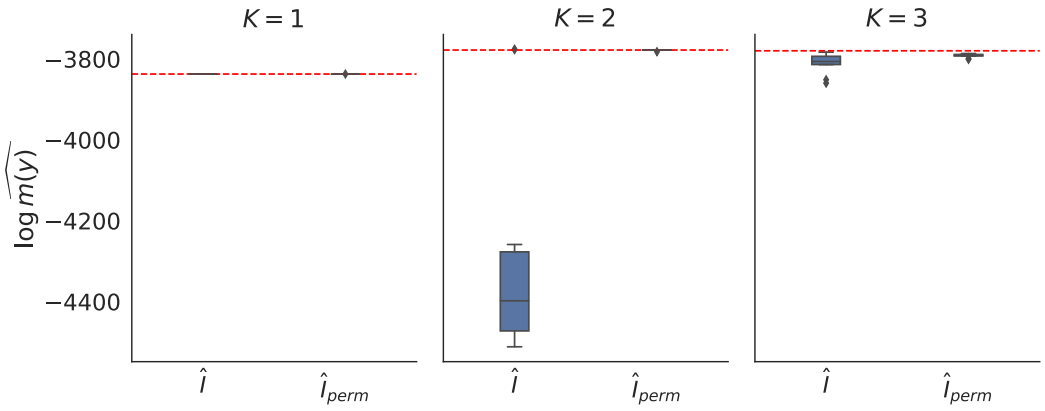


Figure 4.2: Example 1 : Boxplots of 10 repetitions of distributed marginal likelihood estimates for different values of K where I is estimated through $\hat{I}$ and $\hat{I}_{perm}$ and the batch marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ are estimated by the permuted Chib's estimator $\hat{m}_{ChibPerm}(\boldsymbol{y}_s)$ (cf Equation (2.22) of Chapter 2). The same number of Gibbs sampling iterations T is used for both methods. For each K, the reference red dashed line is computed on the whole dataset $\boldsymbol{y}$ using $\hat{m}_{ChibPerm}(\boldsymbol{y})$.

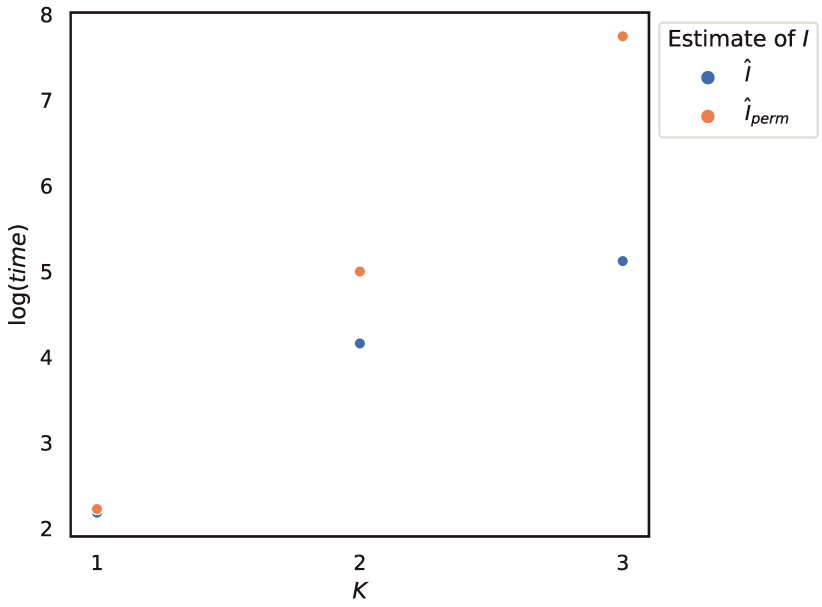


Figure 4.3: Example 1 : Log of the average time required to compute one repetition of the boxplots in Figure 4.2, for $K = 1, 2$ and 3.

The reasons why $\hat{I}_{perm}$ is an improved version of $\hat{I}$ compare easily to the reasons why the fully permuted Chib's estimator introduced in Chapter 2 outperforms the standard Chib's estimator. Intuitively, although $\hat{I}$ is theoretically unbiased, in practice its performance relies on a very good mixing behavior of the underlying S MCMC algorithms targeting the sub-posterior distributions, and in particular a thorough and balanced exploration of each $K!$ modal configurations, both within and across the MCMC samplers. This is an unrealistic expectation in general.

Notice that in the extreme case where only one modal configuration $\boldsymbol{\sigma}$ is visited by the S MCMC samplers, such as in the example of Figure 2.4 in Chapter 2, it is likely that for each $t = 1, \ldots, T$, all but one of the $K!^{S-1}$ integrals in $\hat{I}_{perm}$ should have a value of about 0. In practice, it is impossible to know if this will be the case nor which of the $K!^{S-1}$ integrals is the one with a non-negligible value.

For a given allocation vector $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_S)$, it is convenient to express estimator $\hat{I}_{perm}$ as the average of a measure of compatibility of the permuted allocation vector $\boldsymbol{z} = (\boldsymbol{z}_1, \sigma_2(\boldsymbol{z}_2), \ldots, \sigma_S(\boldsymbol{z}_S))$ over all possible combinations of permutations $\boldsymbol{\sigma} = (\sigma_2, \ldots, \sigma_S) \in \mathfrak{S}_K^{S-1}$. That is,

$$\hat{I}_{perm} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{K!^{S-1}} \sum_{\sigma_2, \ldots, \sigma_S \in \mathfrak{S}_K} \chi(\boldsymbol{z}^{(t)}; \sigma_2, \ldots, \sigma_S)$$

where

$$\chi(\boldsymbol{z}^{(t)}; \sigma_2, \ldots, \sigma_S) := \int \tilde{\pi}(\boldsymbol{\vartheta} | \boldsymbol{z}_1^{(t)}, \boldsymbol{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\boldsymbol{\vartheta} | \sigma_s(\boldsymbol{z}_s^{(t)}), \boldsymbol{y}_s) d\boldsymbol{\vartheta}.$$

Since not all $K!^{S-1}$ configurations $\boldsymbol{\sigma} \in \mathfrak{S}_K^{S-1}$ should make a significant contribution to $\hat{I}_{perm}$ then, if one is able to sample configurations from a well chosen

distribution on $\mathfrak{S}_K^{S-1}$, an importance sampling strategy could be derived and reduce significantly the computational cost of $\hat{I}_{perm}$.

4.4 Importance Sampling estimate of I

Based on the intuition that most of the $K!^{S-1}$ integrals at each iteration t are not relevant for evaluating $\hat{I}_{perm}$, we derive in this section an Importance Sampling (IS) strategy to reduce the explosive computational cost of $\hat{I}_{perm}$.

The problem of constructing a probability distribution over the discrete space $\mathfrak{S}_K^{S-1}$ that puts high probability on configurations $\boldsymbol{\sigma}$ that results in large values of $\chi(\boldsymbol{z};\boldsymbol{\sigma})$ is not straightforward. Indeed, it is not possible to determine a priori such configurations without computing $\chi(\boldsymbol{z};\boldsymbol{\sigma})$ for all $\boldsymbol{\sigma} \in \mathfrak{S}_K^{S-1}$. However, assume that $\boldsymbol{\vartheta} \sim \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{z}_1, \boldsymbol{y}_1)$, then it is reasonable to assume that $\prod_{i:z_{si}=l} p(y_{si}|\vartheta_k)$ is a good approximation to the probability that cluster l of batch s and cluster k of batch 1 'match'. The general idea of our approach is to use the values of the parameters $\boldsymbol{\vartheta}$ sampled from the reference sub-posterior as *anchor* points that will help reconstruct a coherent labeling of the clusters across the batches. Therefore, define for each $s = 2, \ldots, S$ the *matching* matrix

$$P_s = \begin{pmatrix} p_{s11} & \cdots & p_{s1K} \\ \vdots & \vdots & \vdots \\ p_{sK1} & \cdots & p_{sKK} \end{pmatrix}$$

where

$$p_{slk} = \prod_{i:z_{si}=l} p(y_{si}|\vartheta_k)$$

for $k, l = 1, \ldots, K$. For a given batch s, the $l-$th row of matrix P_s gives the matching probabilities of cluster l of batch s with each cluster of batch 1.

Then consider the discrete probability distribution defined for all $\sigma \in \mathfrak{S}_K$ by

$$q_{\sigma_s}(\sigma) \propto \prod_{k=1}^{K} p_{sk\sigma(k)} \tag{4.6}$$

that measures the probability that each cluster k of batch s is matched with cluster $\sigma(k)$ of batch 1. An importance distribution on the configurations $(\sigma_2, \ldots, \sigma_S) \in \mathfrak{S}_K^{S-1}$ can then be defined as

$$q_{\boldsymbol{\sigma}}(\sigma_2, \ldots, \sigma_S) = \prod_{s=2}^{S} q_{\sigma_s}(\sigma_s) \tag{4.7}$$

for all $\boldsymbol{\sigma} = (\sigma_2, \ldots, \sigma_S) \in \mathfrak{S}_K^{S-1}$.

Several remarks can be made about the probability distribution (4.7) above. First, it assumes that the choice of the permutations σ_s is made independently for all $s = 2, \ldots, S$. Hence, simulating a proposal configuration $\boldsymbol{\sigma}$ from $q_{\boldsymbol{\sigma}}$ is equivalent to sampling for each $s = 2, \ldots, S$ a permutation from q_{σ_s} which has a computational complexity of about $\mathcal{O}(K!)$ once the matrix P_s is computed. Second, the definition

q_{σ_s} in (4.6) implies that two different clusters in batch s cannot be matched to the same cluster in the reference batch. Finally, any other batch than batch 1 can be chosen as reference. In practice, it is possible that at an iteration t the allocation vector of $\boldsymbol{z}_s^{(t)}$ only induces $K_s^+ < K$ non-empty clusters. At each iteration, it is then recommended to choose as the reference batch the one that has the largest number of non-empty clusters.

From a computational point of view, due to the discreteness of the importance distribution $q_{\boldsymbol{\sigma}}$, it is sensible to choose the number M of importance simulations at each iteration t according to the number of configurations $\boldsymbol{\sigma}$ with non-negligible weights w.r.t $q_{\boldsymbol{\sigma}}$. For each s, define

$$ESS_s = \frac{\left(\sum_{\sigma \in \mathfrak{S}_K} q_{\sigma_s}(\sigma)\right)^2}{\sum_{\sigma \in \mathfrak{S}_K} q_{\sigma_s}(\sigma)^2} \tag{4.8}$$

which is a measure of the number of permutations with non-negligible q_{σ_s} probability. Notice that in the particular case where $q_{\sigma_s}(\sigma) \propto 1$ for all permutations σ, then $ESS_s = K!$, whereas if only one $q_{\sigma_s}(\sigma)$ has non-zero value, then $ESS_s = 1$. Then, at each iteration t, a sensible choice for $M^{(t)}$ is given by, for example,

$$M^{(t)} = \prod_{s=2}^{S} \lceil ESS_s \rceil .$$

This whole procedure has a considerably reduced computational cost compared to $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$. At each iteration t, each column of P_s requires n/S likelihood evaluations, which makes a total cost of about $O(Kn/S)$ for evaluating P_s. Note that computing the matrices P_s, for $s = 2, \ldots, S$ should be done within each worker separately and in parallel. Then, computing the $K!$ weights of the discrete importance distribution q_{σ_s} simply requires $K!$ simple operations. Finally, sampling from the global discrete importance distribution is done in $M^{(t)}$ basic operations. This adds up to a global cost of $O\Big(T(Kn/S + K! + \bar{M})\Big)$ operations compared to $O\Big(TK!^{S-1}\Big)$, where $\bar{M}$ denotes the maximum number of importance simulations.

In addition, notice that this importance sampling strategy satisfies the desirable requirement that data is not shared across the workers, the only communication being the parameters sampled from the reference sub-posterior distribution.

The above strategy yields the following estimator $\hat{I}_{IS}$, which is unbiased for I by Proposition 4.9. The whole procedure for deriving this estimator is summarized in Algorithm 16.

Proposition 4.7. *Let* $\{\boldsymbol{z}_s^{(t)}\}_{t=1}^T \sim \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$ *for* $s = 1, \ldots, S$. *Then*

$$\hat{I}_{IS} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \frac{1}{M^{(t)}} \sum_{m=1}^{M^{(t)}} \frac{1}{q_{\boldsymbol{\sigma}}(\sigma_2^{(t,m)}, \ldots, \sigma_S^{(t,m)})} \chi(\boldsymbol{z}^{(t)}; \sigma_2^{(t,m)}, \ldots, \sigma_S^{(t,m)}) \tag{4.9}$$

where for all t, $\{(\sigma_2^{(t,m)}, \ldots, \sigma_S^{(t,m)})\}_{m=1}^{M^{(t)}} \overset{i.i.d}{\sim} \pi_{\boldsymbol{\sigma}}$, *is an unbiased estimator for* I.

Proof. For all $t = 1, \ldots, T$,

$$\mathbb{E}\left[\frac{1}{M^{(t)}}\sum_{m=1}^{M^{(t)}}\frac{1}{\pi_{\boldsymbol{\sigma}}(\sigma_1^{(t,m)},\ldots,\sigma_S^{(t,m)})}\chi(\boldsymbol{z}^{(t)};\sigma_2^{(t,m)},\ldots,\sigma_S^{(t,m)})\middle|\boldsymbol{z}^{(t)}\right] = \sum_{\sigma_1,\ldots,\sigma_S\in\mathfrak{S}_K}\chi(\boldsymbol{z}^{(t)};\sigma_2,\ldots,\sigma_S)$$

Hence,

$$\mathbb{E}\left[\hat{I}_{IS}\right] = \mathbb{E}\left[\frac{1}{TK!^{S-1}}\sum_{t=1}^{T}\sum_{\sigma_1,\ldots,\sigma_S\in\mathfrak{S}_K}\chi(\boldsymbol{z}^{(t)};\sigma_2,\ldots,\sigma_S)\right] = \mathbb{E}\left[\hat{I}_{perm}\right] = I$$

□

Algorithm 16 : Importance sampling for distributed marginal likelihood computation

Input : For all $s = 1, \ldots, S$, $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_{t=1}^T$ samples from a MCMC sampler with stationary distribution $\pi(\boldsymbol{\vartheta}, \boldsymbol{z}_s|\boldsymbol{y}_s)$, possibly run in parallel.

Do in parallel
  Estimate $\tilde{m}(\boldsymbol{y}_s)$, possibly using the sample $\{\boldsymbol{\vartheta}_s^{(t)}, \boldsymbol{z}_s^{(t)}\}_t$, for $s = 1, \ldots, S$;
end
for $t = 1, \ldots, T$ **do**
  Do in parallel
    Compute the p.m.f of q_{σ_s} for all s using (4.6);
  end
  Choose $M^{(t)}$ adaptively for using ESS_s defined in (4.8);
  Draw $M^{(t)}$ configurations $\{\sigma_2^{t,m}, \ldots, \sigma_S^{(t,m)}\}_{m=1}^M$ from $q_{\boldsymbol{\sigma}}$;
end
Compute $\hat{I}_{IS}$ using (4.9);
Return $m_{\hat{I}_{S,IS}}(\boldsymbol{y}) := Z^S \prod_{s=1}^S \widehat{\tilde{m}(\boldsymbol{y}_s)} \times \hat{I}_{IS}$;

Example 2. In this example, we generate $n = 1000$ *i.i.d* observation from the standard normal distribution. The choice of the prior distribution and prior hyper-parameters is identical to the setting of Example 1. Figure 4.4 indicates that $\hat{I}_{IS}$ is a good approximation to $\hat{I}_{perm}$ with a considerably reduced computational time, in particular for $K \geq 3$, as shown by Figure 4.5. However, despite being very good at approximating $\hat{I}_{perm}$, it seems that the latter suffers from a pathological variance leading to a downward bias in the log scale. A better understanding of the situation at hand can be provided by the variance of $\hat{I}$ given in Buchholz et al. 2022 as

$$\mathbb{V}(\hat{I}) = \frac{I^2}{T}\mathbb{V}_{\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})}\left(\frac{\pi(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})}{\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})}\right)$$

where as before $\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S}) = \prod_{s=1}^S \tilde{\pi}(\boldsymbol{z}_s|\boldsymbol{y}_s)$.
This expression indicates that if the tails of $\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})$ happen to be thinner than

that of $\pi(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})$, then it is possible that the variance of $\hat{I}$ be infinite. As intuited by Buchholz et al. 2022, as the number of splits S increases, it should be expected that the approximation of the full posterior given by $\tilde{\pi}(\boldsymbol{z}_{1:S}|\boldsymbol{y}_{1:S})$ worsens. In practice, it is difficult to understand the behavior of this discrete distribution and how it reacts to an increase in S and/or K, in our model specification. This discrepancy between $\tilde{\pi}$ and π could be at the origin of the observed better performance of $\hat{I}_{perm}$ and $\hat{I}_{IS}$ in the well-specified scenario ($K = 2$) compared to the ill-specified cases. In any case, this observation calls for a new approach to the problem and, in particular, the derivation of a new identity with hopefully better theoretical properties when confronted to mixture models.

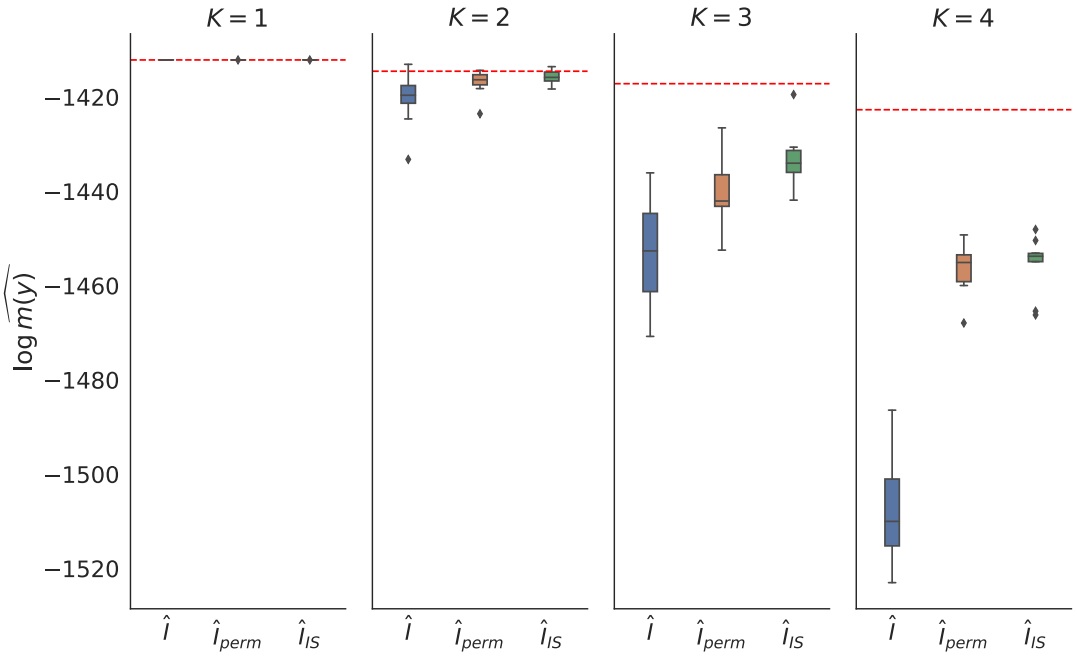


Figure 4.4: Example 2 : Boxplots of 10 repetitions of distributed marginal likelihood estimates for different values of K where I is estimated through $\hat{I}$, $\hat{I}_{perm}$ and $\hat{I}_{IS}$ and the batch marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ are estimated by the permuted Chib's estimator $\hat{m}_{ChibPerm}(\boldsymbol{y}_s)$ (cf Equation (2.22) of Chapter 2). The same number of Gibbs sampling iterations T is used for all methods. For each K, the reference red dashed line is computed on the whole dataset $\boldsymbol{y}$ using $\hat{m}_{ChibPerm}(\boldsymbol{y})$.

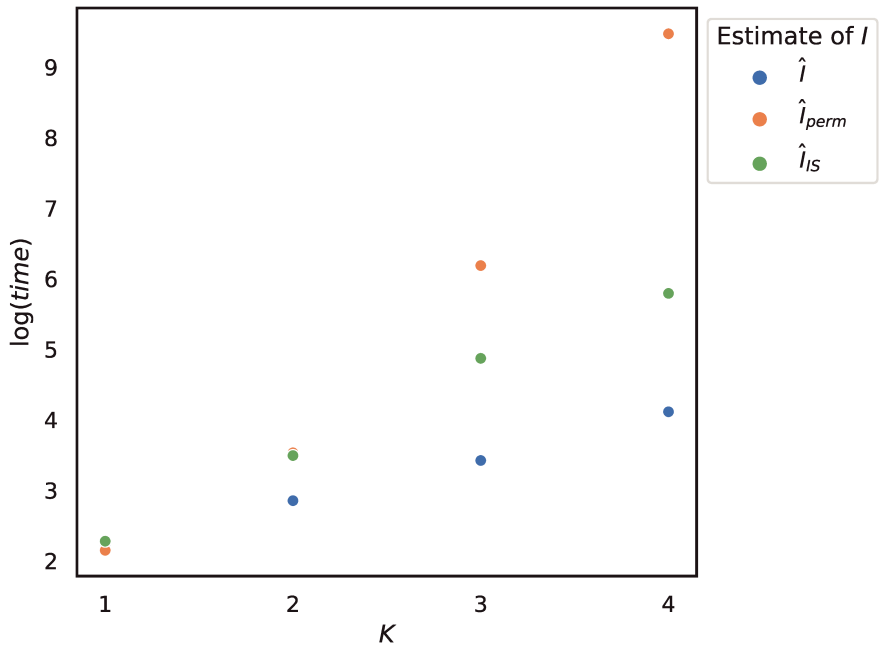


Figure 4.5: Example 1 : Log of the average time required to compute one repetition of the boxplots in Figure 4.4, for $K = 1, 2, 3$ and 4.

4.5 A Sequential Monte Carlo strategy

Another way to tackle the issue of estimating I is to realize that it simply is the normalizing constant of the distribution proportional to the product of the sub-posterior distributions. Define for all $s = 1, \ldots, S$, the distribution proportional to the product of the first s sub-posteriors

$$\tilde{\pi}_s(\boldsymbol{\vartheta}) = \frac{\prod_{l=1}^{s} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_l)}{Z_s} \tag{4.10}$$

where

$$Z_s = \int \prod_{l=1}^{s} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_l).$$

It is interesting to remark that integral I simply is the normalizing constant Z_S of distribution $\tilde{\pi}_S(\boldsymbol{\vartheta})$. Given the complex nature of $\tilde{\pi}_S(\boldsymbol{\vartheta})$, it is expected that standard Monte-Carlo methods should fail at estimating Z_S. However, this new representation of I can lead to a modified identity linking the full marginal likelihood $m(\boldsymbol{y})$ to the successive product distributions $\tilde{\pi}_s(\boldsymbol{\vartheta})$ which is derived below.

Proposition 4.8 (A new identity)**.** *For some data $\boldsymbol{y}$ and a statistical model for which the likelihood function factorizes as $p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{s=1}^{S} p(\boldsymbol{y}_s|\boldsymbol{\vartheta})$, the marginal likelihood of the data can be decomposed as*

$$m(\boldsymbol{y}) = Z^S \times \tilde{m}(\boldsymbol{y}_1) \times \prod_{s=2}^{S} \int \pi_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$$

Proof.

$$\begin{aligned}
m(\boldsymbol{y}) &= Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta} \\
&= Z^S \prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s) Z_S \\
&= Z^S \left[\prod_{s=1}^{S} \tilde{m}(\boldsymbol{y}_s)\right] \left[\prod_{s=2}^{S} \frac{Z_s}{Z_{s-1}}\right] \\
&= Z^S \times \tilde{m}(\boldsymbol{y}_1) \prod_{s=2}^{S} \tilde{m}(\boldsymbol{y}_s) \frac{Z_s}{Z_{s-1}} \\
&= Z^S \times \tilde{m}(\boldsymbol{y}_1) \prod_{s=2}^{S} \frac{\tilde{m}(\boldsymbol{y}_s)}{Z_{s-1}} \int \left[\prod_{l=1}^{s-1} \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_l)\right] \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s) d\boldsymbol{\vartheta} \\
&= Z^S \times \tilde{m}(\boldsymbol{y}_1) \prod_{s=2}^{S} \int \tilde{\pi}_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}
\end{aligned}$$

The third equality is a consequence of the fact that $Z_1 = \int \tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_1) d\boldsymbol{\vartheta} = 1$. □

The advantages of the new identity derived in Proposition 4.8 are twofold. First, it allows to bypass the estimation of all sub-posterior marginal likelihoods by reducing the problem to the estimation of $\tilde{m}(\boldsymbol{y}_1)$ only. Note that the choice of the first batch as a reference in the statement of Proposition 4.8 is completely arbitrary and that any other batch can be used instead. Second, one should notice that the product of integrals on the left-hand side of equation 4.8 can be regarded as the product of the expected value of the unnormalized sub-posterior distributions with respect to the successive product densities $\tilde{\pi}_s(\boldsymbol{\vartheta})$. That is,

$$\prod_{s=2}^{S} \int \tilde{\pi}_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \prod_{s=2}^{S} \mathbb{E}_{\tilde{\pi}_{s-1}(\boldsymbol{\vartheta})} \left[p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta})\right]$$

This representation calls for a (sequential) importance sampling strategy making use of the successive distributions $\tilde{\pi}_s(\boldsymbol{\vartheta})$, $s = 1, \ldots, S$ as importance distributions. More precisely, suppose that at step $s-1$, $\{\boldsymbol{\vartheta}^{(m)}\}_{m=1}^{M}$ is distributed according to $\tilde{\pi}_{s-1}(\boldsymbol{\vartheta})$, then compute the importance weights proportional to the ratio $\tilde{\pi}_s(\boldsymbol{\vartheta}^{(m)})/\tilde{\pi}_{s-1}(\boldsymbol{\vartheta}^{(m)})$,

$$w_s^{(m)} = \frac{\prod_{l=1}^{s} \left[p(\boldsymbol{y}_l|\boldsymbol{\vartheta}^{(m)}) \tilde{\pi}(\boldsymbol{\vartheta}^{(m)})\right]}{\prod_{j=1}^{s-1} \left[p(\boldsymbol{y}_j|\boldsymbol{\vartheta}^{(m)}) \tilde{\pi}(\boldsymbol{\vartheta}^{(m)})\right]} = p(\boldsymbol{y}_s|\boldsymbol{\vartheta}^{(m)}) \tilde{\pi}(\boldsymbol{\vartheta}^{(m)}). \tag{4.11}$$

Then notice that

$$\frac{1}{T} \sum_{t=1}^{T} w_s^{(m)} \approx \int \tilde{\pi}_{s-1}(\boldsymbol{\vartheta}) p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \mathbb{E}_{\tilde{\pi}_{s-1}(\boldsymbol{\vartheta})} \left[p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta})\right].$$

The particles $\{\boldsymbol{\vartheta}^{(m)}\}_m$ are then usually resampled according to their normalized weights $W_s^{(m)} \propto w_s^{(m)}$ and to avoid the potential particle degeneracy phenomenon

typical of Sequential Monte Carlo samplers, $\{\boldsymbol{\vartheta}^{(m)}\}_m$ should be *moved* according to a π_s-invariant MCMC kernel. Assuming samples from the sub-posterior distributions $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$ were collected in parallel with a Gibbs sampler, we propose a Metropolis Hasting step using as a proposal the sub-posterior distribution $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$. That is, at iteration s, for particle $\boldsymbol{\vartheta}^{(m)}$, propose a new particle $\boldsymbol{\vartheta}'$ from the MCMC sample with invariant distribution $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$. Then compute the Metropolis Hasting acceptance probability,

$$\begin{aligned}
\alpha(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}^{(m)}) &= \frac{\tilde{\pi}_s(\boldsymbol{\vartheta}')}{\tilde{\pi}_s(\boldsymbol{\vartheta}^{(m)})} \frac{\tilde{\pi}(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_s)}{\tilde{\pi}(\boldsymbol{\vartheta}'|\boldsymbol{y}_s)} \\
&= \frac{\prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}'|\boldsymbol{y}_s)}{\prod_{s=1}^{S} \tilde{\pi}(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_s)} \frac{\tilde{\pi}(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_s)}{\tilde{\pi}(\boldsymbol{\vartheta}'|\boldsymbol{y}_s)} \\
&= \frac{\prod_{s=1}^{S-1} \tilde{\pi}(\boldsymbol{\vartheta}'|\boldsymbol{y}_s)}{\prod_{s=1}^{S-1} \tilde{\pi}(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_s)} \\
&= \frac{\prod_{s=1}^{S-1} p(\boldsymbol{\vartheta}'|\boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{\vartheta}')}{\prod_{s=1}^{S-1} p(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{\vartheta}^{(m)})}. \qquad (4.12)
\end{aligned}$$

Finally, at step S, compute the estimate

$$\prod_{s=2}^{S} \frac{1}{M} \sum_{m=1}^{T} w_s^{(m)} \approx \prod_{s=2}^{S} \mathbb{E}_{\pi_s(\boldsymbol{\vartheta})} \left[p(\boldsymbol{y}_s|\boldsymbol{\vartheta}) \tilde{\pi}(\boldsymbol{\vartheta}) \right]$$

which is to be plugged into 4.8 to obtain an estimate $\hat{m}_{S,SMC}(\boldsymbol{y})$ of $m(\boldsymbol{y})$. Note that computing the Metropolis acceptance ratio (4.12) can be done in parallel, without sharing any data across the workers. Indeed, at each iteration $s = 2, \ldots, S$ of the SMC sampler, the candidates $\boldsymbol{\vartheta}'$ are chosen in advance among the MCMC sample $\{\boldsymbol{\vartheta}_s^{(t)}\}_{t=1}^{T}$ targeting $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$. Then, the set of candidates is sent to each worker which in turn returns all the ratios constituting (4.12). The above procedure is summarized in Algorithm 17 below.

Algorithm 17 : SMC for distributed marginal likelihood computation

Input : For all $s = 1, \ldots, S$, $\{\boldsymbol{\vartheta}_s^{(t)}\}_{t=1}^T$ samples from a MCMC sampler with stationary distribution $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$, possibly run in parallel.

Do in Worker 1
  Estimate $\tilde{m}(\boldsymbol{y}_1)$, possibly using the sample $\{\boldsymbol{\vartheta}_1^{(t)}\}_t$;
  Initialize the particles $\{\boldsymbol{\vartheta}^{(m)}\}_m^M$ using a subsample of size M from the sample $\{\boldsymbol{\vartheta}_1^{(t)}\}_t$;
  Compute $w_1^{(m)} = p(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_1)\tilde{\pi}(\boldsymbol{\vartheta}^{(m)})$;
end
for $s = 2, \ldots, S$ **do**
  `/* Reweight */`
  Compute the weights $w_s^{(m)}$ of the particles $\{\boldsymbol{\vartheta}^{(m)}\}$ using (4.11) in Worker s;
  `/* Resample */`
  Resample the particles multinomially according to their weights $w_s^{(m)}$;
  `/* Move */`
  Choose M candidates $\{\boldsymbol{\vartheta}'^{(m)}\}_m^M$ among the sample $\{\boldsymbol{\vartheta}_s^{(t)}\}_t^T$;
  Do in parallel
    Compute the ratio $p(\boldsymbol{\vartheta}'^{(m)}|\boldsymbol{y}_s)\tilde{\pi}(\boldsymbol{\vartheta}'^{(m)})/p(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{y}_1)\tilde{\pi}(\boldsymbol{\vartheta}^{(m)})$
  end
  for $m = 1, \ldots, M$ **do**
    Compute the Metropolis acceptance ratio α given in (4.12);
    Accept $\boldsymbol{\vartheta}'^{(m)}$ with probability α;
  end
end
Return $\hat{m}_{S,SMC}(\boldsymbol{y}) = Z^S \widehat{\tilde{m}(\boldsymbol{y}_1)} \prod_{s=2}^S \frac{1}{M} \sum_{m=1}^M w_s^{(m)}$;

Example 2. ***(continued)*** Reusing the same data as in the previous example, we add to our analysis the distributed SMC marginal likelihood estimator derived in Algorithm 17. Once again, the data $\boldsymbol{y}$ is splitted into $S = 3$ batches and the distributed marginal likelihood estimates are displayed on Figure 4.6. Note that all the considered algorithms require as a starting point S samples of size T from the sub-posterior distributions $\tilde{\pi}(\boldsymbol{\vartheta}|\boldsymbol{y}_s)$. For fair comparison, the same value of T is chosen for all methods. More precisely, we choose $T = (K-1)10000 + 1000$ for all values of K considered in this experiment. The number of particles M chosen for the SMC algorithm is set to $(K-1)5000 + 200$. Note that the estimate of the marginal likelihood based on estimator $\hat{I}_{perm}$ is not included for $K = 5$ due to its prohibitive computational cost of $O(5!^2 T)$.

It can be seen that SMC clearly outperforms the other approaches based either on $\hat{I}$, $\hat{I}_{IS}$ or $\hat{I}_{perm}$ and this no matter the value of K. The resulting marginal likelihood estimates consistently fall much closer to the reference value given by ChibPerm with $S = 1$ (red dashed line). Moreover, the required computational time to obtain the SMC estimates is comparable to the other algorithms for $K \leq 2$, but clearly much lower for greater values of K.

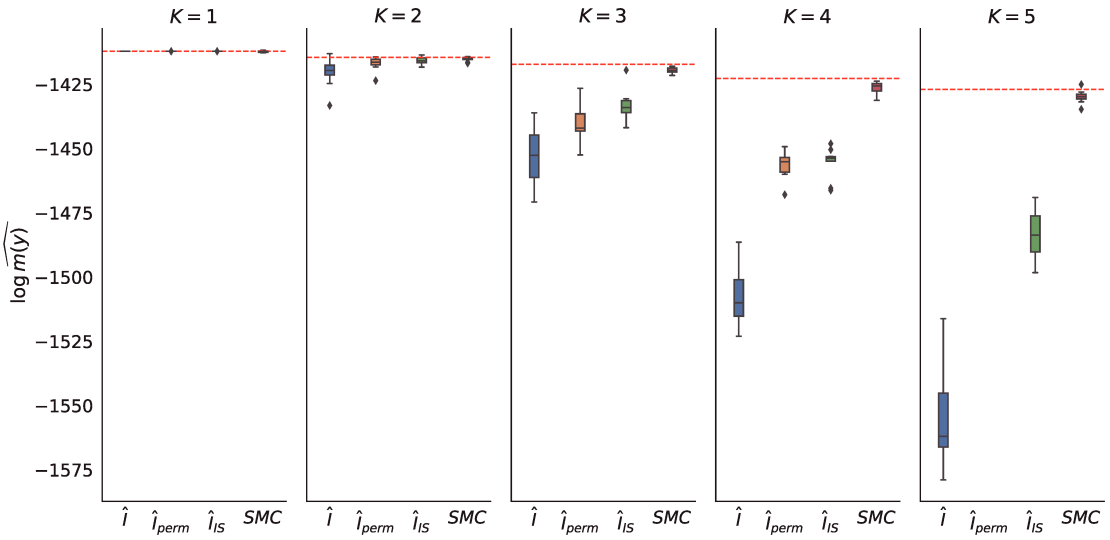


Figure 4.6: Example 2 : Boxplots of 10 repetitions of the distributed marginal likelihood estimated through $\hat{I}$ and $\hat{I}_{perm}$, $\hat{I}$ and SMC for different values of K where and the batch marginal likelihoods $\tilde{m}(\boldsymbol{y}_s)$ are estimated by the permuted Chib's estimator $\hat{m}_{ChibPerm}(\boldsymbol{y}_s)$ (cf Equation (2.22) of Chapter 2). The same number of Gibbs sampling iterations T is used for both methods. For each K, the reference red dashed line is computed on the whole dataset ($S = 1$) $\boldsymbol{y}$ using $\hat{m}_{ChibPerm}(\boldsymbol{y})$.

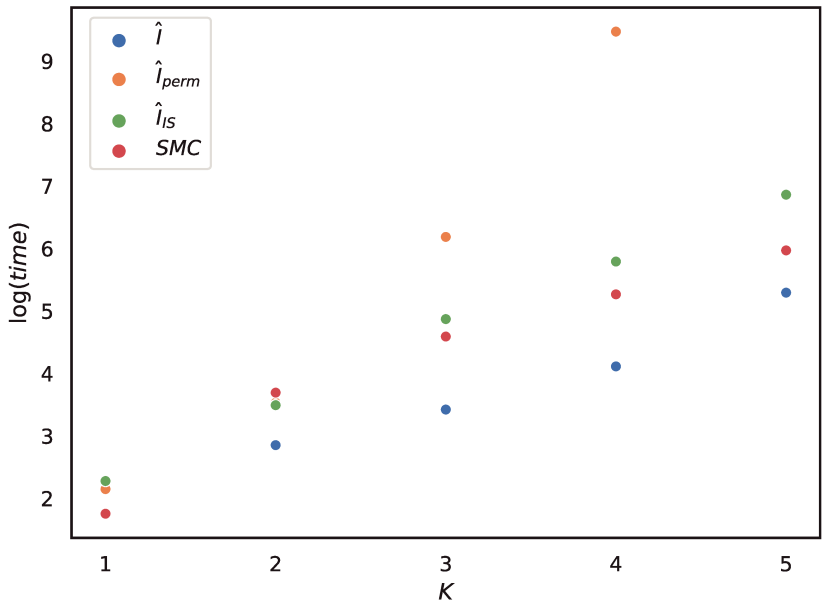


Figure 4.7: Example 2 : Log of the average time required to compute one repetition of the boxplots in Figure 4.6, for $K = 1, 2, 3, 4$ and 5.

4.6 Simulation Study

4.6.1 Experiment 1. The effect of the number of splits S

The aim of this experiment is to assess the distributed marginal likelihood estimates based on $\hat{I}_{IS}$, $\hat{I}_{perm}$ and SMC on a moderately large data set of $n = 20000$ observations and to study the effect of the number of splits S both on the computing time, and on the Monte-Carlo variance of the estimates. To do so, we sample data $\boldsymbol{y}$ from the 2-component mixture

$$0.5\mathcal{N}(2,1) + 0.5\mathcal{N}(5,1)$$

which histogram is plotted on Figure 4.8.

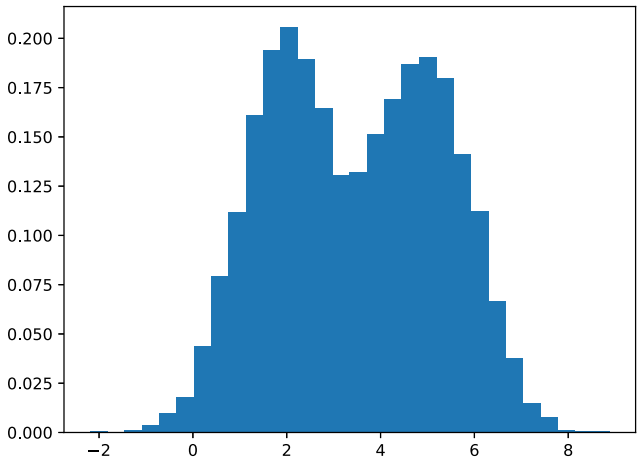


Figure 4.8: Histogram of the data used for Experiment 1. $n = 20000$.

For $S = 1, \ldots, 10$, we successively fit a mixture of K Normal distributions for $K = 2, \ldots, 5$. We choose the conditionally conjugate Normal-Inverse Gamma prior. That is, for all $k = 1, \ldots, K$, we assume $\sigma_k^2 \sim \Gamma^{-1}(a, b)$ and $\mu_k|\sigma_k^2 \sim \mathcal{N}(\mu_0, \sigma_k^2/\lambda)$ where Γ^{-1} is the inverse gamma distribution in the shape and scale parametrization. The hyperparameters (a, b, μ, λ) are derived empirically following recommendations from Raftery 1996 : $b = 0.36(\overline{y^2} - \overline{y}^2), \mu_0 = \bar{y}, 1/\lambda = (y_{max} - y_{min})/2.6$. Finally the prior on the mixture weights is chosen to be Dirichlet with concentration parameter $\boldsymbol{\alpha} = \mathbf{1}_K$. Note that this choice of prior ensures that $\pi(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{z}) := \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2|\boldsymbol{y}, \boldsymbol{z})$ is available in closed form, which is a prerequisite for the tractability of estimators $\hat{I}_{IS}$ and $\hat{I}_{perm}$. The choice of hyperparameter a, however, is constrained by the number of splits S since it must be greater than $(3/2)(S-1)$, as given by Proposition 4.2. In order to assess the effect of S on the estimation of the marginal likelihood, it is not desirable to choose one value of a for each scenario. Therefore, we choose $a = 15$ so that it is compatible with the greatest value of S that we consider in this experiment, namely 10. The number of Gibbs sampling iterations T targeting the sub-posterior distributions $\tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{z}_s|\boldsymbol{y}_s)$ of the Gibbs sampler, a prerequisite to all

the methods considered, is chosen to depend solely on K so as to isolate the effect of S. We choose $T = 50000(K-1) + 1000$ with a burn in of 5%. Whenever the estimation of the marginal likelihood of a batch $\tilde{m}(\boldsymbol{y}_s)$ is necessary, the permuted Chib's estimator ($\hat{m}_{ChibPerm}(\boldsymbol{y})$) is used. In order to assess the precision of our estimates, the scenario $S = 1$ is represented as a red dashed line, computed with the same number T of Gibbs sampling iterations. Indeed, note that all three estimator $\hat{m}_{S,\hat{I}_{IS}}$, $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$ and $\hat{m}_{S,SMC}(\boldsymbol{y})$ reduce to $\hat{m}_{ChibPerm}(\boldsymbol{y})$ whenever $S = 1$.

Figure 4.9 shows the resulting distributed marginal likelihood estimators for each S and K. The estimator $m_{S,\hat{I}_{perm}}(\boldsymbol{y})$ is not always displayed as its computational cost $O\left(TK!^{S-1}\right)$ is too high for some of the values of S and K considered. Note that each repetition used to construct the boxplots uses a different partitioning of the data $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)$ into S batches, so that the observed variation is mostly due to the Monte Carlo variance.

For $K = 2$, i.e the well-specified scenario, we see that estimators $\hat{m}_{S,\hat{I}_{IS}}(\boldsymbol{y})$ and $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$ show rather good performance despite a slight increase in variance can be observed for larger number of splits. The SMC estimate, despite showing a larger variance, does not seem too much affected by the number of splits.

For values of $K > 2$ however, $\hat{m}_{S,SMC}(\boldsymbol{y})$ is consistently closer to the true value showed by the reference line. The other estimates, on the other hand, show a pathological variance. It is interesting to observe that increasing the number of splits tends to introduce some variance in the estimates. Intuitively, this phenomenon can be understood as the price paid for recombining an increasing number of marginal likelihood estimates through I, or in the case of SMC, converting samples from the sub-posterior distributions into the distribution $\pi_s(\boldsymbol{\vartheta})$ (4.10). This observed variance, however, is much smaller for $\hat{m}_{S,SMC}(\boldsymbol{y})$. Furthermore, as in Example 2, we observe that while $\hat{m}_{S,\hat{I}_{IS}}(\boldsymbol{y})$ is a good approximation to $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$, it does not correct the intrinsic pathological variance of the latter.

The estimated Bayes Factors on Figure 4.10 show that SMC consistently leads to the selection of the appropriate number of components ($K_0 = 2$). Moreover, its approximation to the reference values of the Bayes Factor estimated with $\hat{m}_{ChibPerm}(\boldsymbol{y})$ ($S = 1$) is rather satisfactory, despite a slight increase in the observed variance when the number of splits increases. It is to be noted that while the Bayes Factors estimated through $\hat{m}_{S,\hat{I}_{IS}}(\boldsymbol{y})$ and $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$ also lead to a consistent selection of the right number of components, their approximation to the reference Bayes Factor is quite poor and quickly deteriorates as S increases.

In terms of computational time, the information conveyed by Figure 4.11 is twofold. First, using SMC always brings a computational advantage compared to $S = 1$ (red dashed line) in all scenarios considered. Its computational time does not seem to explode with the number of splits. This is not the case for $\hat{m}_{S,\hat{I}_{IS}}(\boldsymbol{y})$ and $\hat{m}_{S,\hat{I}_{perm}}(\boldsymbol{y})$. Second, for all methods, there seems to be an optimal number of splits beyond which the computational time stops decreasing. This value does not seem to depend on K so it is reasonable to assume that it is a function of the sample size n. This clearly highlights a trade-off between the computational time and the number of splits. On the one hand, increasing S clearly reduces the completion time of the parallel MCMC algorithms. On the other hand, the recombination step gets more costly as the number of estimates to merge increases. Note however that SMC seems

to have a recombination step more robust to an increase in S. In fact, looking at the implementation of Algorithm 17, it can easily be seen that its cost is linear in S.

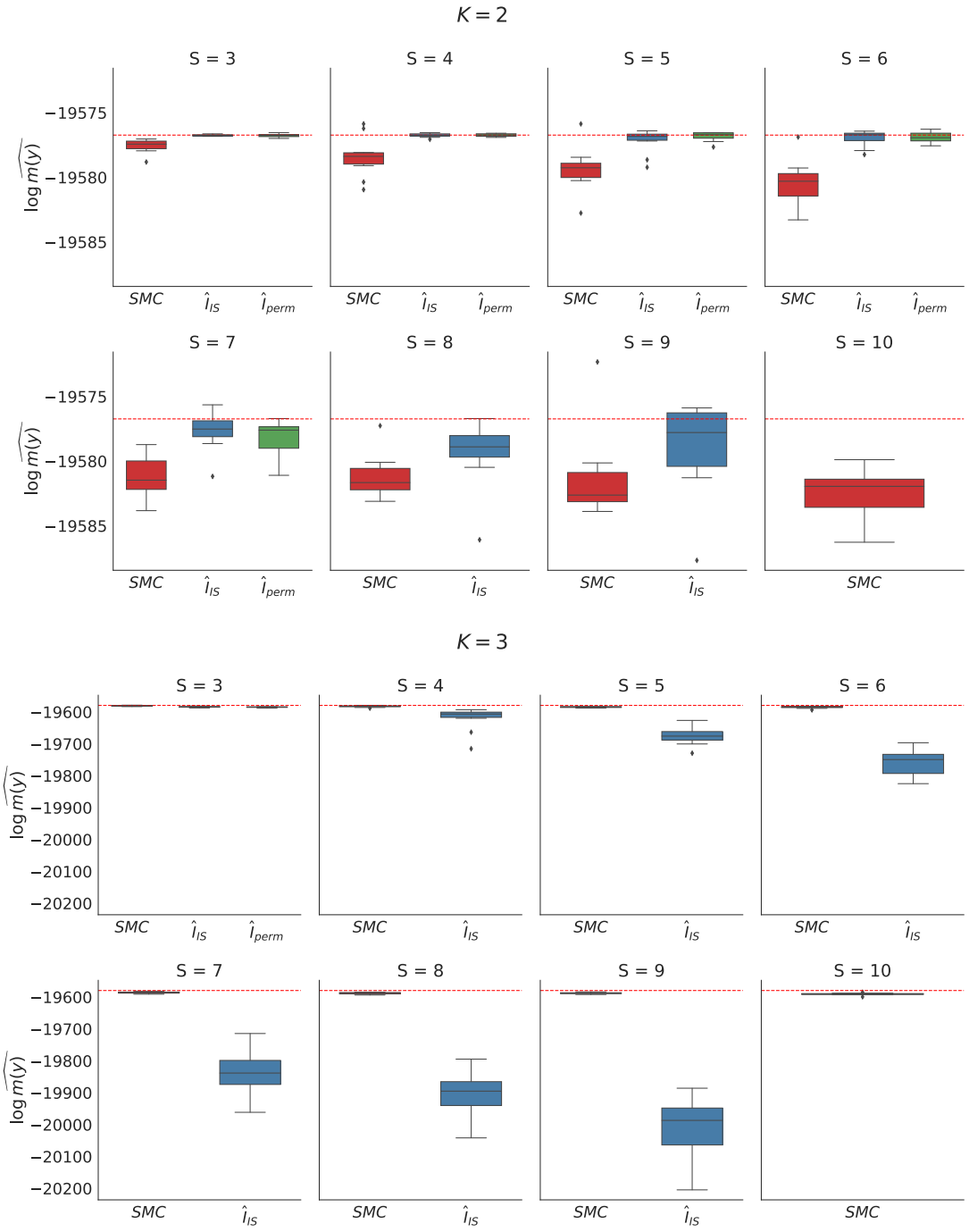


Figure 4.9: Experiment 1. Boxplots for the different methods considered, 10 repetitions each.

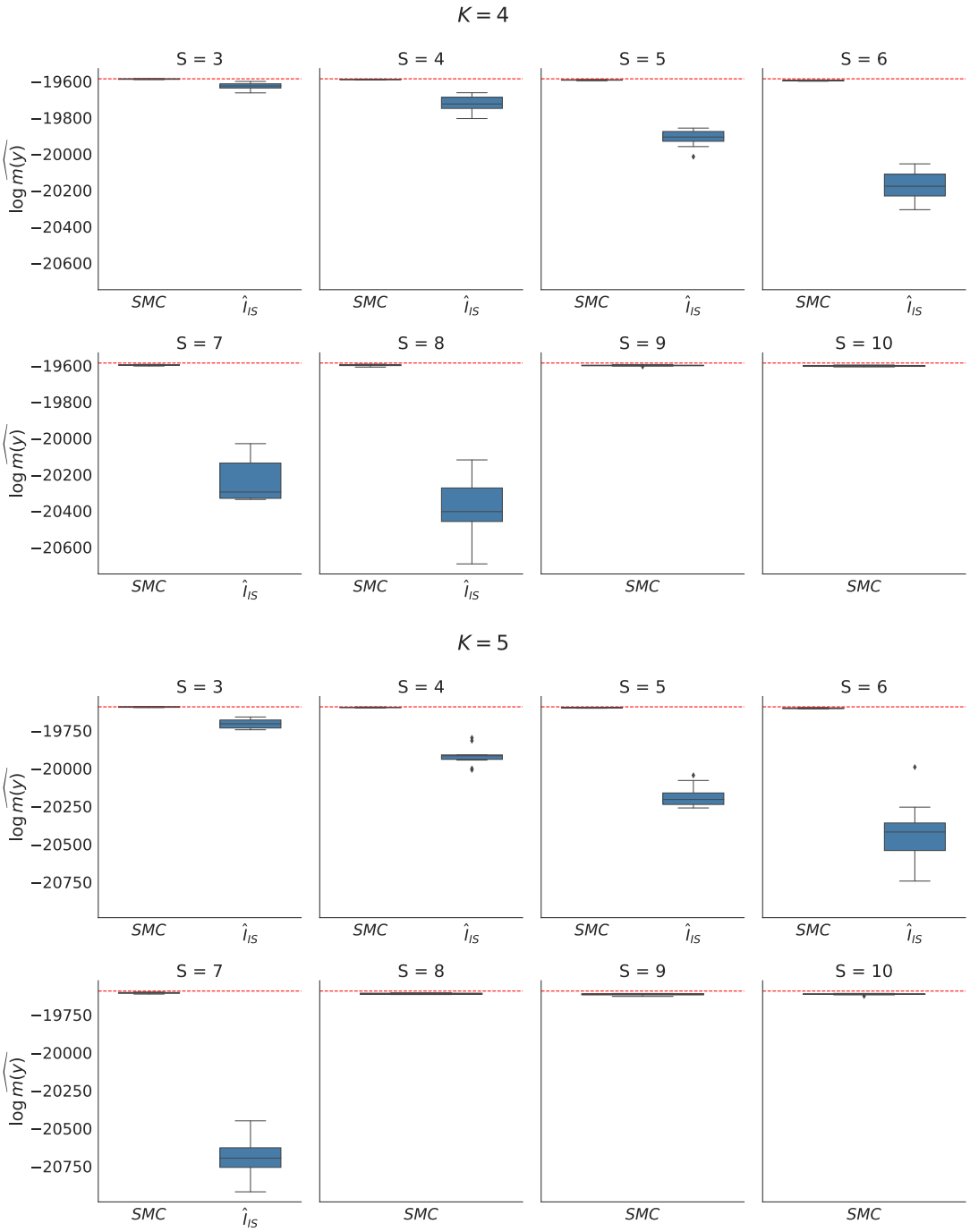


Figure 4.9: Experiment 1 (continued). Boxplots for the different methods considered, 10 repetitions each.

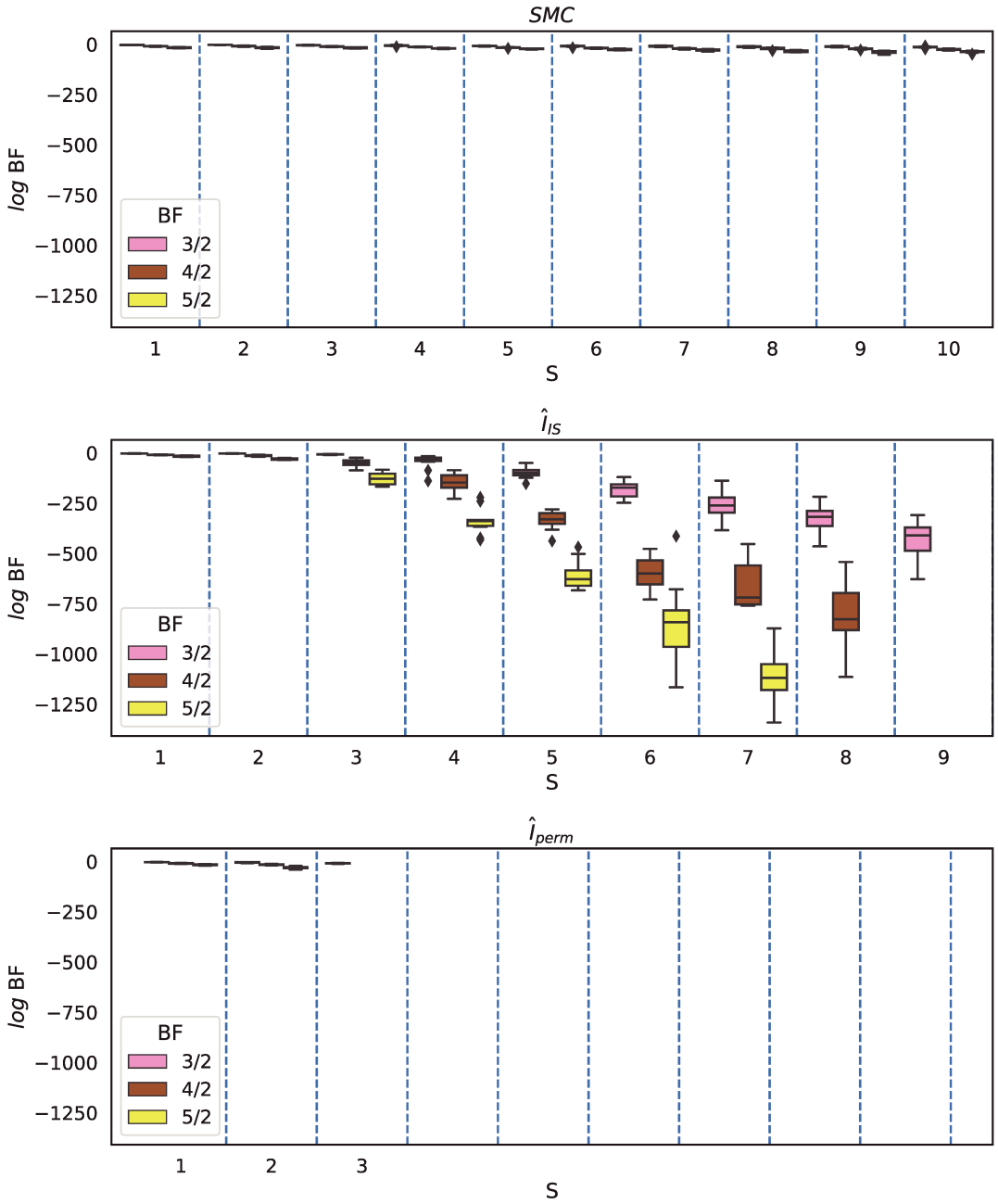


Figure 4.10: Experiment 1 : Log of the Bayes Factor computed with each algorithm. The value for $S = 1$ is computed with $\hat{m}_{ChibPerm}(\boldsymbol{y})$ on the whole dataset. The red dashed line is at $y = 0$ and indicates the $\log BF$ decision threshold.

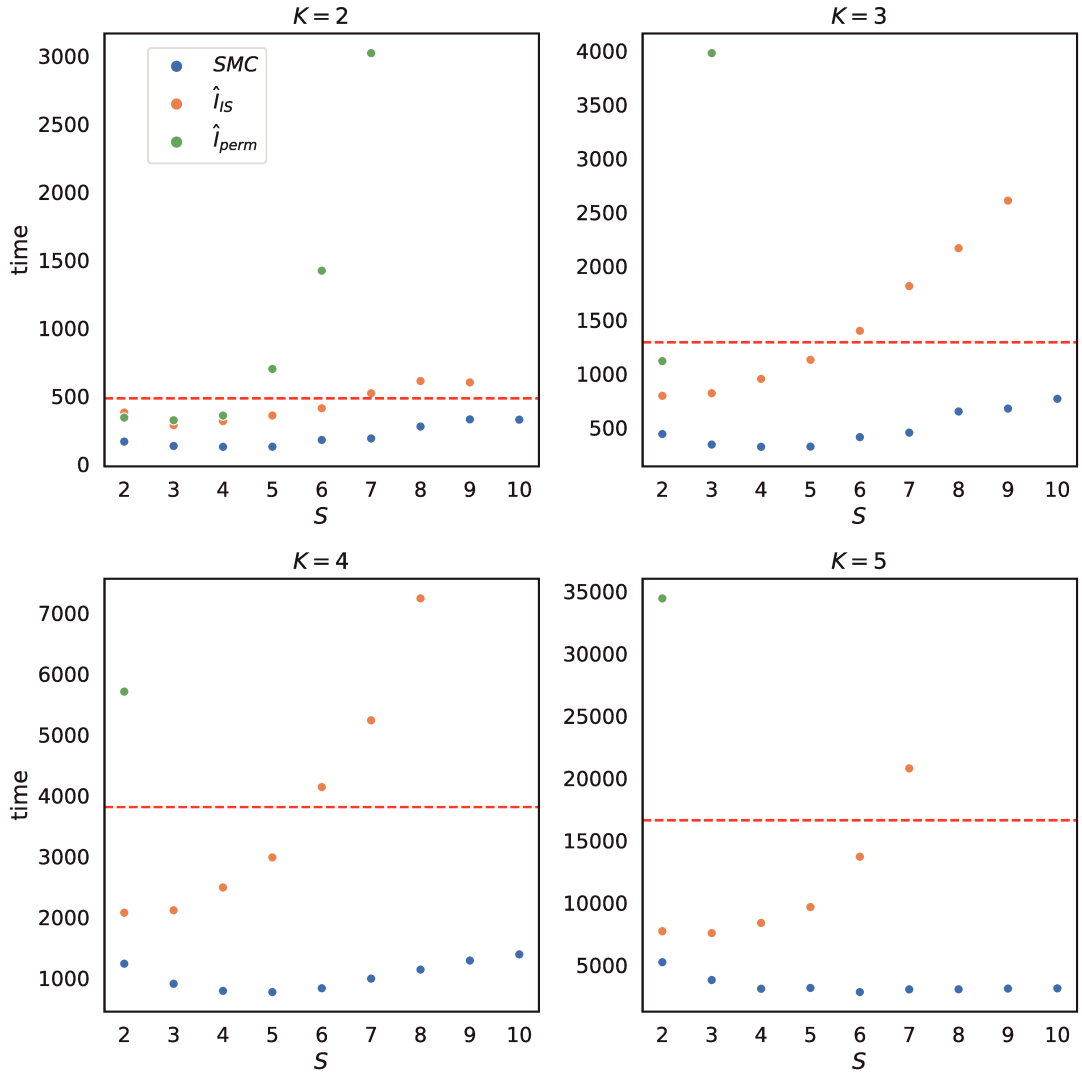


Figure 4.11: Experiment 1 : Log of the average time required to compute one repetition of the boxplots in Figure 4.9, for $K = 1, 2, 3, 4$ and 5.

4.7 Conclusion and perspectives

In this Chapter, we noticed that the distributed marginal likelihood estimation procedure given in Buchholz et al. 2022 is not readily applicable to finite mixture models due to the poor mixing of MCMC samplers caused by their lack of identifiability and highly multimodal posterior. We provided an adapted version of their estimation strategy which directly addresses these issues, and a subsequent version with a more reasonable computational time. Lastly, we derived a new identity bridging the gap between the full marginal likelihood and the batch marginal likelihoods that can be estimated straightforwardly using a Sequential Monte Carlo approach. Besides showing much enhanced performance in the empirical experiments considered, this estimator happens to be valid for all kind of parametric models and relaxes hypotheses such as conjugacy, for instance.

An interesting research avenue would be to find new ways to perform the MCMC step inside the SMC procedure that still preserve the requirement that no data can be shared across the batches. Studying the behavior of this estimator for other models than mixtures could also be interesting. Given its very promising results in such complex scenarios, we are hopeful it could greatly ease the task of Bayesian model selection in the context of tall data for all kind of models.

4.A Appendix

4.A.1 Proof of Proposition 4.2

Proof. Notice that

$$\begin{aligned}\pi(\boldsymbol{\varpi}, \boldsymbol{\theta})^{1/S} &= \pi(\boldsymbol{\varpi})^{1/S} \times \pi(\boldsymbol{\theta})^{1/S} \\ &= \pi(\boldsymbol{\varpi})^{1/S} \times \prod_{k=1}^{K} \pi(\boldsymbol{\theta}_k)^{1/S}.\end{aligned}$$

Now,

$$\pi(\boldsymbol{\varpi})^{1/S} \propto \prod_{k=1}^{K} \pi_k^{(\alpha_k-1)/S} = \prod_{k=1}^{K} \pi_k^{(\alpha_k-1+S)/S-1}$$

hence $\pi(\boldsymbol{\varpi})^{1/S}$ is Dirichlet with parameters $(\alpha_k - 1 + S)/S$, which are always positive for $S \geq 1$.
Moreover,

$$\begin{aligned}\pi(\boldsymbol{\theta}_k)^{1/S} &\propto \left(\frac{1}{\sigma_k^2}\right)^{(a+3/2)/S} \exp\left\{-\frac{2b + \lambda(\mu_k - \mu_0)^2}{S \times 2\sigma^2}\right\} \\ &\propto \left(\frac{1}{\sigma_k^2}\right)^{(a+3/2-(3/2)S)/S+3/2} \exp\left\{-\frac{2(b/S) + (\lambda/S)(\mu_k - \mu_0)^2}{2\sigma^2}\right\}\end{aligned}$$

hence $\pi(\boldsymbol{\theta}_k)^{1/S}$ is still Normal-Inverse Gamma with hyper-parameters as given above, provided $(a + 3/2 - (3/2)S)/S > 0 \Leftrightarrow a > (3/2)(S-1)$. □

4.A.2 Distribution of the product of the augmented sub-posteriors

Proposition 4.9. *For a $K-$component mixture of Normal distributions with prior distribution on the component weights and parameters $(\boldsymbol{\varpi}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = ((\mu_1, \sigma_1^2), \ldots, (\mu_K, \sigma_K^2))$ given by*

$$\Pi(d\boldsymbol{\varpi}, d\boldsymbol{\theta}) = \mathcal{D}(d\boldsymbol{\varpi}|\boldsymbol{\alpha}) \times \prod_{k=1}^{K} \mathcal{NIG}(d\theta_k|\mu_0, \lambda, a, b)$$

then

$$\prod_{s=1}^{S} \tilde{\Pi}(d\boldsymbol{\varpi}, d\boldsymbol{\theta}|\boldsymbol{z}_s, \boldsymbol{y}_s) \propto \mathcal{D}(d\boldsymbol{\varpi}|\tilde{\boldsymbol{\alpha}}) \times \prod_{k=1}^{K} \mathcal{NIG}(d\theta_k|\tilde{\mu}_{0k}, \tilde{\lambda}_k, \tilde{a}_k, \tilde{b}_k)$$

where for all $k = 1, \ldots, K$,

$$\begin{cases} \tilde{\alpha}_k = \sum_{s=1}^{S} \alpha_{sk} \\ \tilde{\mu}_{0k} = \left(\sum_{s=1}^{S} \lambda_{sk} \mu_{0sk} \right) / \sum_{s=1}^{S} \lambda_{sk} \\ \tilde{\lambda}_k = \sum_{s=1}^{S} \lambda_{sk} \\ \tilde{a}_k = \sum_{s=1}^{S} a_{sk} + (3/2)(S-1) \\ \tilde{b}_k = \sum_{s=1}^{S} b_{sk} + (1/2) \left(\sum_{s=1}^{S} \lambda_{sk} \mu_{0sk}^2 - \left(\sum_{s=1}^{S} \lambda_{sk} \mu_{0sk} \right)^2 / \sum_{s=1}^{S} \lambda_{sk} \right) \end{cases}$$

and $(\alpha_{sk}, \mu_{0sk}, \lambda_{sk}, a_{sk}, b_{sk})$ *are the sub-posterior hyper parameters derived in Corollary 4.4.*

Proof. The result is obtained by direct calculation. □

Bibliography

Aiyer, A., Pyun, K., Huang, Y.-Z., O'Brien, D. B., and Gray, R. M. (2005). "Lloyd clustering of Gauss mixture models for image compression and classification". In: *Signal Processing: Image Communication* 20.5, pp. 459–485.

Aldous, D. J. (1985). "Exchangeability and related topics". In: *École d'été de probabilités de Saint-Flour, XIII—1983*. Vol. 1117. Lecture Notes in Math. Springer, Berlin, pp. 1–198.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The Annals of Statistics* 2, pp. 1152–1174.

Argiento, R., Guglielmi, A., and Pievatolo, A. (2010). "Bayesian density estimation and model selection using nonparametric hierarchical mixtures". In: *Computational Statistics & Data Analysis* 54.4, pp. 816–832.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023). "Clustering consistency with Dirichlet process mixtures". In: *Biometrika* 110.2, pp. 551–558.

Basu, S. and Chib, S. (2003). "Marginal likelihood and Bayes factors for Dirichlet process mixture models". In: *Journal of the American Statistical Association* 98.461, pp. 224–235.

Berkhof, J., Van Mechelen, I., and Gelman, A. (2003). "A Bayesian approach to the selection and testing of mixture models". In: *Statistica Sinica* 13.2, pp. 423–442.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Buchholz, A., Ahfock, D., and Richardson, S. (2022). "Distributed Computation for Marginal Likelihood based Model Choice". In: *Bayesian Analysis* 18.2, pp. 607–638.

Buchholz, A., Chopin, N., and Jacob, P. E. (2021). "Adaptive tuning of Hamiltonian Monte Carlo within sequential Monte Carlo". In: *Bayesian Analysis* 16.3, pp. 1–27.

Cai, D., Campbell, T., and Broderick, T. (2021). "Finite mixture models do not reliably learn the number of components". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 1158–1169.

Carvalho, C. M., Lopes, H. F., Polson, N. G., and Taddy, M. A. (2010). "Particle learning for general mixtures". In: *Bayesian Analysis* 5.4, pp. 709–740.

Castillo, I. (2008). "Lower bounds for posterior rates with Gaussian process priors". In: *Electronic Journal of Statistics* 2, pp. 1281–1299.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019). "Model selection for mixture models—perspectives and strategies". In: *Handbook of mixture analysis*.

Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL, pp. 117–154.

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and Inferential Difficulties with Mixture Posterior Distributions". In: *Journal of the American Statistical Association* 95.451, pp. 957–970.

Chen, J. H. (1995). "Optimal rate of convergence for finite mixture models". In: *The Annals of Statistics* 23.1, pp. 221–233.

Chen, M.-H. and Shao, Q.-M. (1997). "On Monte Carlo methods for estimating ratios of normalizing constants". In: *The Annals of Statistics* 25.4, pp. 1563–1594.

Chen, R. and Liu, J. S. (1996). "Predictive updating methods with application to Bayesian classification". In: *Journal of the Royal Statistical Society. Series B. Methodological* 58.2, pp. 397–415.

Chen, Z. and Doss, H. (2019). "Inference for the number of topics in the latent Dirichlet allocation model via Bayesian mixture modeling". In: *Journal of Computational and Graphical Statistics* 28.3, pp. 567–585.

Chérief-Abdellatif, B.-E. and Alquier, P. (2018). "Consistency of variational Bayes inference for estimation and model selection in mixtures". In: *Electronic Journal of Statistics* 12.2, pp. 2995–3035.

Chib, S. (1995). "Marginal likelihood from the Gibbs output". In: *Journal of the American Statistical Association* 90.432, pp. 1313–1321.

Chib, S. and Kuffner, T. A. (July 2016). "Bayes factor consistency". Preprint at arXiv:1607.00292.

Chopin, N. (2002). "A sequential particle filter method for static models". In: *Biometrika* 89.3, pp. 539–552.

Crépet, A. and Tressou, J. (2011). "Bayesian nonparametric model with clustering individual co-exposure to pesticides found in the French diet". In: *Bayesian Analysis* 6.1, pp. 127–144.

Dacunha-Castelle, D. and Gassiat, E. (1997). "The estimation of the order of a mixture model". In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 3.3, pp. 279–299.

Dang, U. J., Gallaugher, M. P. B., Browne, R. P., and McNicholas, P. D. (2023). "Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions". In: *Journal of Classification* 40.1, pp. 145–167.

Dass, S. C. and Lee, J. (2004). "A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives". In: *Journal of Statistical Planning and Inference* 119.1, pp. 143–152.

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 68.3, pp. 411–436.

Diebolt, J. and Robert, C. P. (1990). "Bayesian estimation of finite mixture distributions, Part ii: Sampling implementation". In: *Technical Report 111 , LSTA, Université Paris VI, Paris.*

Drton, M. and Plummer, M. (2017). "A Bayesian information criterion for singular models". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 79.2, pp. 323–380.

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures". In: *Journal of the American Statistical Association* 90.430, pp. 577–588.

Feigin, P. D. and Tweedie, R. L. (1989). "Linear functionals and Markov chains associated with Dirichlet processes". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 105.3, pp. 579–585.

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions". In: *Recent Advances in Statistics*. Elsevier, pp. 287–302.

— (1973). "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* 1, pp. 209–230.

Fong, E. and Holmes, C. C. (2020). "On the marginal likelihood and cross-validation". In: *Biometrika* 107.2, pp. 489–496.

Frühwirth-Schnatter, S. (2004). "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques". In: *The Econometrics Journal* 7.1, pp. 143–167.

— (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

— (2019). "Keeping the balance—Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and Markov mixture models". In: *Brazilian Journal of Probability and Statistics* 33.4, pp. 706–733.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized mixtures of finite mixtures and telescoping sampling". In: *Bayesian Analysis* 16.4, pp. 1279–1307.

Gassiat, E. and Van Handel, R. (2014). "The local geometry of finite mixtures". In: *Transactions of the American Mathematical Society* 366.2, pp. 1047–1072.

Geary, D. N. (1989). "Mixture Models: Inference and Applications to Clustering". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 152.1, pp. 126–127.

Geyer, C. J. (1994). "Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo". In: *Technical Report, School of Statistics, University of Minnesota* 568.

Ghosal, S. (2010). "The Dirichlet process, related priors and posterior asymptotics". In: *Bayesian nonparametrics*. Vol. 28. Camb. Ser. Stat. Probab. Math. Cambridge Univ. Press, Cambridge, pp. 35–79.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). "Convergence rates of posterior distributions". In: *The Annals of Statistics* 28.2, pp. 500–531.

Ghosal, S., Lember, J., and Vaart, A. van der (2008). "Nonparametric Bayesian model selection and averaging". In: *Electronic Journal of Statistics* 2.none, pp. 63–89.

Ghosal, S. and van der Vaart, A. (2007). "Posterior convergence rates of Dirichlet mixtures at smooth densities". In: *The Annals of Statistics* 35.2, pp. 697–723.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Ed. by Springer.

Graham, R. L., Knuth, D. E., Patashnik, O., and Liu, S. (1989). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, pp. 258–259.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, pp. 711–732.

Griffin, J. (2017). "Sequential Monte Carlo methods for normalized random measure with independent increments mixtures". In: *Statistics and Computing* 27.1, pp. 131–145.

Gunawan, D., Dang, K.-D., Quiroz, M., Kohn, R., and Tran, M.-N. (2020). "Subsampling sequential Monte Carlo for static Bayesian models". In: *Statistics and Computing* 30.6, pp. 1741–1758.

Hallock, H. P., Marshall, S. E., Hoen, P. A. '. t., Nygård, J., Hoorne, B., Fox, C., et al. (2021). "Federated Networks for Distributed Analysis of Health Data". In: *Frontiers in Public Health* 9.

Heckman, J. J., Robb, R., and Walker, J. R. (1990). "Testing the Mixture of Exponentials Hypothesis and Estimating the Mixing Distribution by the Methods of Moments". In: *Journal of the American Statistical Association* 85.410, pp. 582–589.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., eds. (2010). *Bayesian nonparametrics*. Vol. 28. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Huang, Z. and Gelman, A. (2005). "Sampling for Bayesian Computation with Large Datasets".

Irwin, M., Cox, N., and Kong, A. (1994). "Sequential imputation for multilocus linkage analysis". In: *Proceedings of the National Academy of Sciences* 91.24, pp. 11684–11688.

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). "Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo". In: *Scandinavian Journal of Statistics* 38.1, pp. 1–22.

Jeffreys, H. S. (1935). "Some Tests of Significance, Treated by the Theory of Probability". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 31, pp. 203–222.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015). "On particle methods for parameter estimation in state-space models". In: *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 30.3, pp. 328–351.

Kong, A., Liu, J. S., and Wong, W. H. (1994). "Sequential imputations and Bayesian missing data problems". In: *Journal of the American Statistical Association* 89.425, pp. 278–288.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). "Adaptive Bayesian density estimation with location-scale mixtures". In: *Electronic Journal of Statistics* 4, pp. 1225–1257.

Lee, J. E. and Robert, C. P. (2016). "Importance sampling schemes for evidence approximation in mixture models". In: *Bayesian Analysis* 11.2, pp. 573–597.

Lenk, P. (2009). "Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods". In: *Journal of Computational and Graphical Statistics* 18.4, pp. 941–960.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (May 2020). "Marginal likelihood computation for model selection and hypothesis testing: an extensive review". In: *SIAM Review, 2022*.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999). "Sequential importance sampling for nonparametric Bayes models: The next generation". In: *Canadian Journal of Statistics* 27.2, pp. 251–267.

Marin, J.-M. and Robert, C. P. (2008). "Approximating the marginal likelihood in mixture models". In: *Bulletin of the Indian Chapter of ISBA* 1, pp. 2–7.

McLachlan, G. (1987). "On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3, pp. 318–324.

McLachlan, G., Lee, S. X., and Rathnayake, S. I. (2019). "Finite mixture models". In: *Annual Review of Statistics and its Application* 6, pp. 355–378.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. Vol. 38. M. Dekker New York.

Mcvinish, R., Rousseau, J., and Mengersen, K. (2009). "Bayesian goodness-of-fit testing with mixtures of triangular distributions". In: *Scandinavian Journal of Statistics* 36.2, pp. 337–354.

Meng, X.-L. and Wong, W. H. (1996). "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration". In: *Statistica Sinica* 6.4, pp. 831–860.

Miller, J. W. and Harrison, M. T. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components". In: *arXiv preprint arXiv:1301.2708*.

— (2018). "Mixture models with a prior on the number of components". In: *Journal of the American Statistical Association* 113.521, pp. 340–356.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (June 2014). "Scalable and Robust Bayesian Inference via the Median Posterior". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1656–1664.

Neal, R. M. (1992). "Bayesian mixture modeling". In: *Maximum Entropy and Bayesian Methods*. Springer, pp. 197–211.

— (1999). "Erroneous results in marginal likelihood from the Gibbs output". Preprint available at https://www.cs.utoronto.ca/ radford/ftp/chib-letter.pdf.

— (2000). "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of Computational and Graphical Statistics* 9.2, pp. 249–265.

Neiswanger, W., Wang, C., and Xing, E. P. (2014). "Asymptotically Exact, Embarrassingly Parallel MCMC". In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*. UAI'14. Quebec City, Quebec, Canada: AUAI Press, pp. 623–632.

Newey, W. K. and West, K. D. (1986). "A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix". In: *Econometrica* 55.3, pp. 703–708.

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 56.1, pp. 3–48.

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models". In: *The Annals of Statistics* 41.1, pp. 370–400.

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture". In: *The Annals of Statistics* 32.5, pp. 2044–2073.

Pajor, A. (2017). "Estimating the marginal likelihood using the arithmetic mean identity". In: *Bayesian Analysis* 12.1, pp. 261–287.

Pearson, K. (1894). "Contributions to the Mathematical Theory of Evolution". In: *Philosophical Transactions of the Royal Society of London. A* 185, pp. 71–110.

Quintana, F. A. and Newton, M. A. (2000). "Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences". In: *Journal of Computational and Graphical Statistics* 9.4, pp. 711–737.

Rabaoui, A., Viandier, N., Duflos, E., Marais, J., and Vanheeghe, P. (2012). "Dirichlet process mixtures for density estimation in dynamic nonlinear modeling: application to GPS positioning in urban canyons". In: *IEEE Transactions on Signal Processing* 60.4, pp. 1638–1655.

Raftery, A. E. (1996). "Hypothesis testing and model selection". In: *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, pp. 163–188.

Ray, S. and Mallick, B. (2006). "Functional clustering by Bayesian wavelet methods". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 68.2, pp. 305–332.

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 59.4, pp. 731–792.

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 73.5, pp. 689–710.

Rousseau, J. and Szabo, B. (2020). "Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors". In: *The Annals of Statistics* 48.4, pp. 2155–2179.

Schäfer, C. and Chopin, N. (2013). "Sequential Monte Carlo on large binary sampling spaces". In: *Statistics and Computing* 23.2, pp. 163–184.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). "Bayes and big data: the consensus Monte Carlo algorithm". In: *International Journal of Management Science and Engineering Management* 11.2, pp. 78–88.

Scricciolo, C. (2014). "Adaptive Bayesian density estimation in Lp-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures". In: *Bayesian Analysis* 9.2, pp. 475–520.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors". In: *Statistica Sinica* 4.2, pp. 639–650.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures". In: *Biometrika* 100.3, pp. 623–640.

Smyth, P. (2000). "Model selection for probabilistic clustering using cross-validated likelihood". In: *Statistics and Computing* 10.1, pp. 63–72.

Srivastava, S., Li, C., and Dunson, D. B. (2018). "Scalable Bayes via barycenter in Wasserstein space". In: *Journal of Machine Learning Research* 19, Paper No. 8, 35.

Tokdar, S. T. and Martin, R. (2021). "Bayesian test of normality versus a Dirichlet process mixture alternative". In: *Sankhya B* 83.1, pp. 66–96.

Ullah, I. and Mengersen, K. (2019). "Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data". In: *Journal of Big Data*.

Verdinelli, I. and Wasserman, L. (1998). "Bayesian goodness-of-fit testing using infinite-dimensional exponential families". In: *The Annals of Statistics* 26.4, pp. 1215–1241.

Yamato, H. (1984). "Characteristic functions of means of distributions chosen from a Dirichlet process". In: *The Annals of Probability* 12.1, pp. 262–267.

Zeng, S., Huang, R., Kang, Z., and Sang, N. (2014). "Image segmentation using spectral clustering of Gaussian mixture models". In: *Neurocomputing* 144, pp. 346–356.

RÉSUMÉ

Ce travail vise à développer de nouveaux outils et procédures pour le problème de la sélection de modèle bayésienne pour les modèles de mélanges. Le facteur de Bayes, défini comme le rapport des vraisemblances marginales calculées pour deux modèles concurrents, est connu pour être consistent dans la plupart des situations. En pratique, l'estimation de la vraisemblance marginale des mélanges finis est une tâche complexe et s'accompagne généralement d'un coût computationnel d'ordre K!, où K est le nombre de composantes du mélange. Nous passons en revue les estimateurs les plus populaires de la vraisemblance marginale pour les mélanges finis et proposons deux méthodes alternatives plus robustes à une augmentation de K et de n, le nombre d'observations. Nous nous intéressons également au modèle de mélange de processus de Dirichlet (DPM) et proposons des estimateurs fiables de la vraisemblance marginale pour de tels modèles non paramétriques. Une application immédiate est la mise en place de tests d'adéquation dans lesquels l'adéquation d'un modèle paramétrique est évaluée par rapport à celle d'une alternative non paramétrique, incarnée par le DPM. Nous montrons que cette procédure est valide en prouvant que le facteur de Bayes est consistent dans ce cadre. Enfin, nous examinons la question de l'estimation distribuée de la vraisemblance marginale pour les mélanges finis, qui reste largement inexplorée jusqu'à présent. Comme c'est généralement le cas avec les mélanges, la plupart des difficultés découlent du manque d'identifiabilité dans l'étiquetage des clusters qu'ils induisent. En utilisant le cadre du Monte Carlo séquentiel, nous développons une méthode robuste qui accélère considérablement le calcul de la vraisemblance marginale en permettant l'échantillonnage selon la loi a posteriori en parallèle.

MOTS CLÉS

Méthodes de Monte Carlo, Bayésien non-paramétrique, mélanges, calcul distribué

ABSTRACT

This PhD dissertation aims at deriving new tools and procedures for the problem of Bayesian model selection for mixture models. The Bayes Factor, defined as the ratio of the marginal likelihood (a.k.a model evidence) computed for two competing models, is known to be consistent in most situations. In practice, the estimation of the marginal likelihood of finite mixtures is a complicated task and usually comes at a computational cost of order K!, where K is the number of mixture components. We review the most popular estimators of the model evidence for finite mixtures and suggest two alternative methods that scale better both with K and n, the number of observations. We also consider the Dirichlet process mixture model (DPM) and derive reliable estimators of the marginal likelihood for such non-parametric models. An immediate application is the derivation of goodness-of-fit tests in which the fit of a parametric model is assessed against that of a non-parametric alternative, embodied by the DPM. We show that this procedure is valid by proving that the Bayes Factor is consistent in this setting. Finally, we consider the issue of distributed model evidence estimation for finite mixtures, which remains largely unexplored so far. As usual with mixtures, most difficulties arise due to the lack of identifiability in their induced cluster labeling. Using the Sequential Monte Carlo framework, we derive a robust method that greatly speeds up evidence computation by allowing posterior sampling to be done in parallel.

KEYWORDS

Monte Carlo methods, Bayesian non-parametrics, mixtures, distributed computing