Advances in Bayesian asymptotics and Bayesian nonparametrics

Caroline Lawless

Jesus College

University of Oxford

A thesis submitted for a degree of

Doctor of Philosophy

October 2023

Statement of originality

I hereby declare that except where specific reference is made to the work of others, the intellectual contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification, or any other university.

Caroline Lawless

October 2023

To my sisters, Anna and Julia

Acknowledgements

Firstly, my deepest gratitude goes to my supervisors Judith, Robin, and Christian, for their guidance and support these past four years. Thank you for your patience, for meeting with me every week, for your ideas, and for giving me the freedom to work on problems that I enjoy. I would especially like to thank Judith and Robin for the precious time they have spent helping me during the final stages of this PhD. I extend my gratitude to Florence and François for agreeing to be my examiners.

An important recognition goes to Julyan who has played a decisive role in this PhD. Thank you for my research internship in 2018, a positive experience that gave me the confidence to do a PhD. Thank you for collaborating with me this year, and for the time you have dedicated to our projects, particularly during my research visit to Grenoble in May and this final week. I would also like to thank Louise, Guillaume, Daria and Filippo for their collaboration and discussions.

I very am grateful to the European Research Council (ERC) who funded this PhD as well as all my my conferences, to Jesus College Oxford for funding my research visit to Grenoble, and to G-Research London further financial support.

One of the highlights of my PhD was the opportunity to work with different teams, in different places, with different people. I would like to thank CEREMADE Paris-Dauphine, and INRIA Grenoble Rhône-Alpes for their warm welcome during my time in Paris and Grenoble. I thank the Department of Statistics: co-members of the Bayesian group, colleagues and friends, for making my time in Oxford so special.

I thank my friends from XVème Athletic Club, and in particular, my coach André, who have been like a second family to me during my time in Paris. I also thank Eric and Françoise Chélin, for their constant support and kindness.

Finally, I thank the most important people in my life: my parents, my sisters Anna and Julia, and my brother Charles, to whom I owe everything. Above all, I thank Morgan, for your support during the most difficult moments, for moving to Oxford with me last year, and for making these last four years great.

Contents

Abstract

Bayesian statistics is a powerful approach to learning real-world phenomena, its strength lying in its ability to quantify uncertainty explicitly by treating unknown quantities of interest as random variables. In this thesis, we consider questions regarding three quite different aspects of Bayesian learning.

Firstly, we consider approximate Bayesian computation (ABC), a computational method suitable for computing approximate posterior distributions for highly complex models, where the likelihood function is intractable but can be simulated from. Previous authors have proved consistency and provided rates of convergence in the case where all summary statistics converge at the same rate as each other. We generalize to the case where summary statistics may converge at different rates, and provide an explicit representation of the shape of the ABC posterior distribution in our general setting. We also show under our general setting that local linear post-processing can lead to significantly faster contraction rates of the pseudo-posterior.

We then focus on the application of Bayesian statistics to natural language processing. The class of context-free grammars, which are standard in the modelling of natural language, have been shown to be too restrictive to fully describe all features of natural language. We propose a Bayesian non-parametric model for the class of 2-multiple context-free grammars, which generalise context-free grammars. Our model is inspired by previously proposed Bayesian models for context-tree garammars and is based on the hierarchical Dirichlet process. We develop a sequential Monte Carlo algorithm to make inference under this model and carry out simulation studies to assess our method.

Finally, we consider some consistency issues related to Bayesian nonparametric mixture models. It has been shown that these models are inconsistent for the number of clusters. In the case of Dirichlet process (DP) mixture models, this problem can be mitigated when a prior is put on the model's concentration hyperparameter α, as is common practice. We prove that Pitman–Yor process (PYP) mixture models (which generalise DP mixture models) remain inconsistent for the number of clusters when a prior is put on α, in the special case where the true number of components in the data generating mechanism is equal to 1 and the discount parameter $\sigma \in (0, 1)$ is a fixed constant. When considering the space over partitions induced by BNP mixture models, point estimators such as the maximum a posteriori (MAP) are commonly used to summarise the posterior clustering structure of such models, which alone can be complex and difficult to interpret. We prove consistency of the MAP partition for DP mixture models when the concentration parameter, α_n, goes deterministically to zero, and when the true partition is made of only one cluster.

Chapter 1

Introduction

In this thesis, I consider various aspects of Bayesian statistics, including Bayesian computation, Bayesian nonparametric modelling, and Bayesian asymptotics. In Section 1.1 I introduce the general context, review some of the relevant literature, and motivate the problems that we will be considering. In Section 1.2 I outline my main contributions.

1.1 Background

In a statistical analysis, data $y_{1:n} = (y_1, \dots, y_n) \in \mathcal{Y}$ is typically assumed to have been generated from some probability distribution $P(\cdot|\theta)$ which can by characterized by a parameter $\theta \in \Theta$, where Θ denotes some parameter space equipped with a metric, d. The goal is to make inference on θ given $y_{1:n}$ i.e. to make conclusions about the underlying probabilistic model describing the data. We will use $Y_{1:n}$ when referring to data as a random variable, and $y_{1:n}$ when referring to a particular observation of data. Throughout, we will assume that P emits a density with respect to some measure, which we denote $p(y|\theta)$.

In this thesis, we consider the Bayesian approach to the problem, which differs from the so-called frequentist approach in that, instead of directly seeking estimates of one "true" parameter θ_0, the inference is based on a probability distribution over Θ. A Bayesian model consists of a prior density $\pi(\theta)$ which describes our prior belief and uncertainty about the parameter θ and a likelihood density $p(y|\theta)$ (as described above) which describes the generative distribution of data given some parameter $\theta \in \Theta$, which can be combined together to form a posterior density $\pi(\theta|y_{1:n})$ using Bayes' rule:

$$\pi(\theta|y_{1:n}) = \frac{p(y_{1:n}|\theta)\pi(\theta)}{\int p(y_{1:n}|\theta)\pi(\theta)d\theta}. \tag{1.1}$$

The Bayesian approach can be easily applied to complex models and, unlike the frequentist approach, provides explicit measures of the uncertainty over the parameter space. For an extensive introduction to Bayesian statistics see Robert et al. (2007).

The first step of a Bayesian analysis consists of choosing an appropriate likelihood density $p(y|\theta)$ on $\mathcal{Y}$ and a suitable prior density $\pi(\theta)$ on the parameter space Θ. If one has some prior belief of where the true parameter is concentrated, this should be incorporated into the prior distribution. Otherwise, a prior distribution with high variance could be a more suitable choice.

One then computes the posterior density $\pi(\theta|y_{1:n})$ using Equation (1.1). Since the likelihood function $p(y_{1:n}|\theta)$ is typically unavailable in closed form, this is often done using computational methods. These include simulation-based methods based on Monte Carlo estimates, and approximate methods, where one designs some pseudo-posterior distribution which is "close" to the true posterior, and from which estimates can more easily be made.

A Bayesian posterior can be summarised by a point estimate, for example, the posterior mean, $E\left(\theta|y_{1:n}\right) = \int \theta d\Pi(\theta|y_{1:n})$ or the maximum a posteriori (MAP) estimator, $\hat{\theta}_n = \arg\max \pi\left(\theta|y_{1:n}\right)$. Point estimates, however, do not make full use of the power of the posterior, and in particular, lack a measure of uncertainty. A more comprehensive representation of the posterior would be to provide either its density (if available), or a large sample of draws from it.

1.1.1 Bayesian asymptotics

Since Bayesian posterior computation is often approximate, before drawing statistical conclusions from them it is crucial to validate their reliability. One way of doing this is by studying their asymptotic properties. We adopt what is called a frequentist-Bayesian point of view and assume the existence of a true parameter θ_0.

A posterior distribution is said to be consistent if its mass concentrates on increasingly small neighbourhoods around the true parameter as the amount of data goes to infinity. Formally, posterior consistency is defined as follows.

Definition 1. *The posterior distribution is said to be consistent at $\theta_0 \in \Theta$ with respect to a metric d on Θ if for any $\epsilon > 0$, the posterior probability of an $\epsilon-$neighbourhood of θ_0, $\mathcal{N}_\epsilon = \{\theta : d(\theta, \theta_0) < \epsilon\}$ converges to 1:*

$$\Pi(\mathcal{N}_\epsilon | y_{1:n}) \to 1$$

in $P(\cdot|\theta_0)-$probability as n goes to infinity.

Consistency is a minimal requirement for a posterior distribution to be considered reliable. A first result of Doob (1949) shows that when d is a metric and (Θ, d) is a complete separable space, the posterior is guaranteed to concentrate on a neighbourhood Θ_0 of θ_0 as long as Θ_0 has strictly positive measure under the prior and as long as θ is identifiable. In other words, the posterior is consistent everywhere except for a set of values having measure zero under the prior. This result is interesting but weak since it fails on a null set which is unknown and which depends on the prior.

In the case of independent and identically distributed data, Schwartz's theorem (Schwartz (1965)) guarantees consistency for finite dimensional models under testing conditions of the model and under the condition that the prior puts enough positive mass around the true parameter (in the sense of Kullback-Leibler divergence). Barron et al. (1999) extend this to the non-iid case.

A more refined asymptotic property than consistency is posterior contraction rates (also known as posterior concentration rates). They provide a measure of how fast the posterior distribution shrinks around the true parameter, and are defined as follows.

Definition 2. *A rate of contraction of the posterior distribution with respect to a metric d on Θ is defined as a sequence $(\epsilon_n)_{n\geq 1}$ such that*

$$\Pi(\theta : d(\theta, \theta_0) \leq M_n \epsilon_n | y_{1:n}) \to 1$$

in $P(\cdot|\theta_0)-$probability as n goes to infinity, where M_n is any monotone increasing sequence. The best possible (i.e. the smallest) sequence $(\epsilon_n)_{n\geq 1}$ satisfying the above is called the optimal rate of contraction.

In their seminal paper, Ghosal et al. (2000) develop a general methodology to obtain posterior contraction rates, which is extended to the case of non iid observations in Ghosal and Van Der Vaart (2007).

Finally, one may be interested in the asymptotic shape of the posterior distribution. It is well-known that in finite-dimensional regular models the MLE $\hat{\theta}_n$ has the following Gaussian limiting distribution in $P(\cdot|\theta_0)-$probability:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, \mathcal{I}_n(\theta_0)^{-1}), \tag{1.2}$$

where $\mathcal{I}_n$ is the Fisher information matrix. Note that the above is a frequentist result: the "true" parameter θ_0 is a fixed constant and randomness is due to the data which is distributed according to the likelihood model. The Bernstein-von Mises Theorem due to Laplace (1810) provides a Bayesian analogue of Equation (1.2) for finite-dimensional models. It states that under some regularity conditions, the posterior distribution has the following Gaussian limit in $\Pi(\cdot|y_{1:n})-$ probability:

$$(\sqrt{n}(\theta - \hat{\theta}_n)|y_{1:n}) \to \mathcal{N}(0, \mathcal{I}_n(\theta_0)^{-1}).$$

Note that the above is a Bayesian result: unlike in Equation (1.2) where randomness comes from the data y_n, the Bernstein-von Mises Theorem considers data $y_{1:n}$ (and thus $\hat{\theta}_n$) to be fixed, and randomness comes from the posterior distribution over the parameter space Θ.

As we will describe in detail in the next section, a common modelling choice in Bayesian statistics involves defining the parameters to be of infinite dimension. Such models are referred to as Bayesian nonparametric (BNP) models. While general results such as Schwartz' theorem and the Bernstein Von Mises theorem are widely applied to finite models, they are not always applicable in nonparametric settings. Providing asymptotic guarantees for Bayesian nonparametric models can be challenging, and is often done in a case by case basis (see, for example Walker and Hjort (2001), Walker (2003), Walker (2004), and Lijoi et al. (2007) for consistency results, and see Rousseau (2016), Ray and van der Vaart (2021) and Franssen and van der Vaart (2022) for Bernstein Von Mises results in BNP settings). General asymptotic results also fail to hold in situations where Bayesian computation is based on estimations of some pseudo-posterior distribution which approximates the true posterior.

1.1.2 Bayesian nonparametrics

A Bayesian nonparametric model is defined as a Bayesian model over the space of infinitely many parameters (Bernardo and Smith (2009)). Bayesian nonparametric models are popular for their flexibility and are particularly useful when dealing with highly complex data.

The most popular Bayesian nonparametric prior is the Dirichlet process (DP), introduced by Ferguson (1973). There are several equivalent representations of the DP. Sethuraman (1994) defines the DP in a constructive way called the "stick breaking" representation as follows.

Definition 3. *If* $V_i \sim_{iid} Beta(1, \alpha)$ *for* $i = 1, \ldots,$ *if* $p_j = V_j \prod_{i=1}^{j-1}(1-V_i)$ *for* $j = 1, 2, \ldots,$

and $\theta_1, \theta_2, \ldots \sim_{iid} Q_0$ then the discrete random probability measure

$$G = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

is distributed according to a Dirichlet process with concentration parameter α and base distribution Q_0. We write $G \sim DP(\alpha, Q_0)$.

Blackwell and MacQueen (1973) provide a characterization of the DP based on the generative distribution of data points drawn from draws from it: if $(\theta_1, \ldots, \theta_n, \theta_{n+1}) \sim G$ and $G \sim DP(\alpha, Q_0)$, then conditional on $(\theta_1, \ldots, \theta_n)$, the $(n+1)^{th}$ observation θ_{n+1} is equal to θ_j with probability $\frac{n_j}{\alpha+n}$ (where n_j represents the number of components in $(\theta_1, \ldots, \theta_n)$ that take the same value as θ_j) and is distributed according to Q_0 with probability $\frac{\alpha}{\alpha+n}$. This process is commonly referred to as the "Chinese restaurant process" due to an analogy of customers sitting at tables in a Chinese restaurant: when a customer (the element θ_{n+1}) enters a restaurant, they sit at the j^{th} table (i.e. take the value θ_j) with probability $\frac{n_j}{\alpha+n}$, and sit at a new unoccupied table (i.e. take a new and unique value) with probability $(\frac{\alpha}{\alpha+n})$.

Although the Dirichlet process itself is a measure on the space of discrete measures, it can be used to model continuous data by convolving it with some kernel. Given a class of kernels $k(y|\theta)$, Lo (1984) defines the Dirichlet process mixture model as a model with density defined by $f(y)$, where

$$\begin{aligned} G &\sim DP(\alpha, Q_0) \\ f(y) &= \int k(y|\theta) G(d\theta). \end{aligned} \tag{1.3}$$

In Dirichlet process mixture models, each observation is associated with one component $\theta \in \Theta$ and then distributed according to $k(y|\theta)$ conditional on θ. DP mixture models are thus perfectly suited to modelling clustered data: two data points belong to the same cluster if they are both associated with the same component $\theta \in \Theta$. Unlike the case of finite mixture models, where the G in (1.3) is a finite discrete measure, they allow for an unbounded number of components. The complexity of the model may grow as the training data becomes available.

Grouped structure in data is common across a broad range of domains, including genetics (Gabriel et al. (2002)) where the groups are the halotypes of binary markers of the human genome, information retrieval (Blei et al. (2003)) where the groups are the topics of a set of documents, or natural language processing (Liang et al. (2007)) where the groups are the parts of speech of a grammar. Teh et al. (2004) extend the DP mixture model further with their hierarchical Dirichlet process (HDP) model, where separate groups are modeled with separate Dirichlet processes. The Dirichlet processes are linked together with a common base distribution, which itself is modeled with another Dirichlet process. Mathematically, their model can be represented as follows.

$$\begin{aligned} &G_0 \sim DP(\alpha_0, Q_0) \\ &G_j \sim DP(\alpha_j, G_0) \qquad \forall j \in J \end{aligned}$$

where J represents the number of groups one wishes to model. If G_0 were some continuous distribution, the probability of different groups sharing components would be zero. On the other hand, if G_0 were a finite discrete measure, the model would be too restrictive. The use of a Dirichlet process as a common base distribution allows both across-group and within-group clustering, without compromising model flexibility. Beal et al. (2001) propose a closely related model to the HDP, which is a hidden Markov model where the transitions are modeled using a HDP (one considers the states of the hidden Markov model to be the groups). Finkel et al. (2007) use an adaptation of the HDP model for three models over trees, with various dependency assumptions among the children at each branch.

While the DP remains the most standard Bayesian nonparametric prior, numerous extensions exist. Pitman–Yor processes (PYPs) are a simple extension of Dirichlet processes, developed by Perman et al. (1992) and further investigated by Pitman (1995) and Pitman and Yor (1997). PYPs introduce an extra parameter $\sigma \in [0, 1)$ (called the discount parameter), that allows for flexible control of the clustering behavior, and can be characterized by a "stick-breaking" construction, identical to that of the Dirichlet process in Definition 3, except with the V_i's distributed as $V_i \sim_{ind} \text{Beta}(1-\sigma, \alpha + i\sigma)$. When G is distributed according to a Pitman–Yor process with concentration parameter α, base distribution Q_0, and with discount parameter σ we write $G \sim PYP(\alpha, \sigma, Q_0)$. The PYP can also be characterized by the generative distribution of data points drawn from draws from it, and by its corresponding Chinese restaurant process analogy: a new customer (the value θ_{n+1}) sits at the j^{th} table (i.e. takes the value θ_j) with probability $\frac{n_j - \sigma}{\alpha + n}$, and sits at a new unoccupied table (i.e. takes a new and unique value) with probability $(\frac{\alpha + n\sigma}{\alpha + n})$. The Dirichlet process mixture model and hierarchical Dirichlet process model described above can be trivially adapted to PYPs by replacing DPs by PYPs.

Beyond the Dirichlet process and the Pitman–Yor process, other nonparametric priors do exist, for instance, the class of Gibbs-type priors (De Blasi et al. (2013)), which naturally generalise DPs and PYPs, but these are beyond the scope of this thesis.

1.1.3 Bayesian modelling for grammars

The goal of natural language processing (NLP) is to develop algorithms that allow computers to understand natural language. NLP applications include speech recognition, translation, and language understanding, among others, and due to the large quantity of linguistic data available on the internet have enjoyed a significant amount of research attention over the last twenty years (Manning and Schutze (1999)).

One object of interest when studying linguistic data are the *parse trees* which describe the structure of each sentence in the language. An example of a parse tree for the English sentence "They solved the problem with Bayesian statistics" is provided in Figure 1.1.

A grammar is what defines the structure of a language. Chomsky (1956) defined a formal grammar $\mathcal{G}$ to be four-tuple $(\mathcal{A}, \mathcal{B}, \mathcal{R}, \mathcal{S})$ where $\mathcal{A}$ is the set of *terminal symbols* (the words in the language), and where $\mathcal{B}, \mathcal{R}$ and $\mathcal{S}$ are related to their underlying structure. In particular, when considering the parse trees of the sentences in a language, the symbol $\mathcal{S}$ is the *start symbol* at every root node, $\mathcal{B}$ is the set of *nonterminal symbols* at internal nodes, and $\mathcal{R}$ is the set of *rules* which define which branching patterns can occur.

Inferring the grammar that best describes some natural language based on a finite set of sentences is a challenging task. We approach this problem by using *probabilistic grammars* which are defined as formal grammars which additionally have some set $\mathcal{J}$ of probabilities, with one probability assigned to each of the rules in $\mathcal{R}$. The probability of any sentence is defined to be the product of the probabilities assigned to each of the rules in that sentence's parse tree.

Chomsky (1956) classified grammars in terms of the complexity of the rules that they allow. In order of increasing complexity he defined the following four classes of grammars.

$$\text{Regular} \subset \text{Context-free} \subset \text{Context-sensitive} \subset \text{Recursively enumerable} \tag{1.4}$$

The more complex a grammar model is the more features of a language it may be able to capture. However, the more complex the grammar model, the more expensive the inference may be in terms of computational time. It is common for natural language to be modeled using context-free grammars, and probabilistic context-free grammars have been a core modelling technique for many aspects of linguistic structure (Charniak (1996), Collins (2003)). All of the rules of a context-free grammar (when in its Chomsky normal form) must be written in one of the following two forms

$$B_j \to B_{k_1} B_{k_2} \tag{1.5}$$

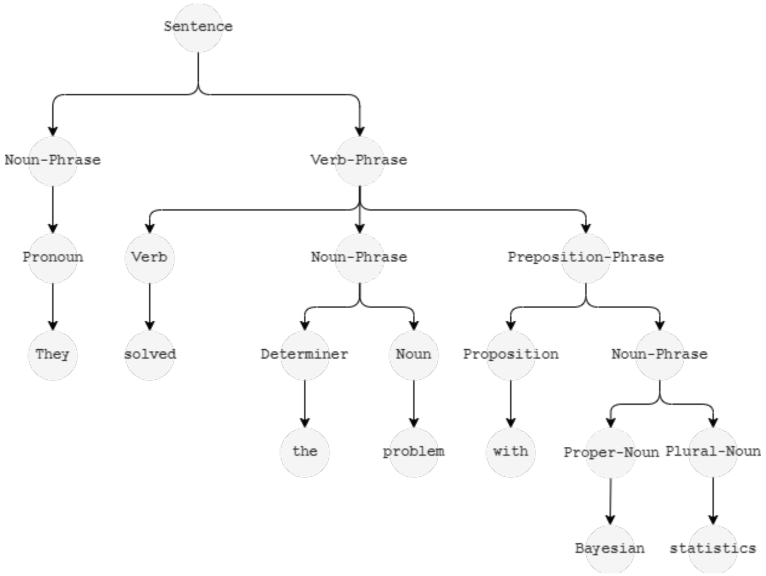


Figure 1.1: A parse tree for the sentence *They solved the problem with Bayesian statistics.*

$$B_j \to a_k \tag{1.6}$$

where $B_j, B_{k_1}, B_{k_2} \in \mathcal{B}$ and where $a_k \in \mathcal{A}$. Figure 1.1 illustrates a sentence and its parse tree generated from a context-free grammar describing the English language. In this grammar, $\mathcal{S}$ is the nonterminal [Sentence], $\mathcal{B}$ is the set of English parts of speech (for example [Noun], [Verb], [Proposition], etc.), and $\mathcal{A}$ is the set of English words. Elements of the set $\mathcal{R}$ include

$$\text{[Noun-Phrase]} \to \text{[Determiner] [Noun]}$$

and

$$\text{[Proper-Noun]} \to \text{“Bayesian”}.$$

Traditionally, estimation of probabilistic context-free grammars has been done using variants of the inside-outside algorithm (Baker (1979), Lari and Young (1990)) which is based on the frequentist technique of expectation maximization (EM) estimation. Despite the fact that the majority of early work in statistical NLP has been non-Bayesian, it can be argued that the Bayesian approach is perfectly suited to NLP. Indeed by an appropriate choice of prior, Bayesian methods can favor sparseness which is typical of linguistic data. Johnson et al. (2007) show that a finite Bayesian model for probabilistic context-free grammars can infer linguistic structure in situations where maximum likelihood methods such as the Inside-Outside algorithm only produce a trivial grammar, using the example of the sparse grammar describing the morphology of the Bantu language Sesotho. Furthermore, by using nonparametric priors, the Bayesian approach allows the number of nonterminal symbols and rules to be learned adaptively with the data, providing flexible models with an unbounded number of latent parameters.

All of the rules in a context-free grammar involve overwriting one nonterminal symbol (the B_j on the left-hand side of Equation (1.5) and Equation (1.6)) with either a pair of nonterminal symbols (as in Equation (1.5)) or with a single terminal symbol (as in Equation (1.6)). The rules of a context-free grammar can thus be modeled as a multi-group mixture: two rules are in the same group if and only if they both overwrite the same nonterminal symbol. Liang et al. (2007) model the different groups of rules separately whilst maintaining a global link across all rules by means of hierarchical Dirichlet processes, in their HDP-PCFG model. Goldwater et al. (2006) extend this to model parse trees for context-free grammars with various different dependency assumptions across child nodes.

Despite the extensive work on natural language inference based on context-free grammars, it is well-known that these models do not capture all features of human natural language. Shieber (1985) demonstrate this for the particular case of the Swiss German language. Since the class of context-sensitive grammars, which comes above the class of context-free grammars in terms of complexity, is considered too complex in practice for simple inference purposes, researchers have proposed intermediate classes of grammars that lie in between context-free and context-sensitive in terms of complexity. Examples of these include head grammars (Pollard (1984)), tree-adjoining grammars (Joshi et al. (1969)), and 2-multiple context-free grammars (Seki et al. (1991)). Until now, no Bayesian model has been proposed for these extensions, and all inference has been frequentist.

1.1.4 BNP mixtures

Mixture models are commonly used in statistical analysis of heterogeneous data where observations are assumed to come from a number of different populations or groups. Due to their flexibility and simplicity, they are popular across a wide range of applications, including healthcare (Ramírez et al. (2019)), econometrics (Frühwirth-Schnatter et al. (2012)), and ecology (Attorre et al. (2020)).

In a mixture model, each observation is assumed to come from exactly one group, and each group is characterized by some density, which usually comes from some parametric family. Mathematically, a mixture model over data $y_{1:n}$ can be characterized by the distribution $F^\star$ with pdf with respect to some measure μ

$$f^\star(y) = \sum_{j=1}^{t} p_j^\star k(y|\phi_j^\star), \qquad t \in \mathbb{N} \tag{1.7}$$

where the $p_j^\star$ are probability weights in $(0, 1)$ summing to one, and where the $k(\cdot|\phi_j^\star)$ are probability kernels, each depending on some parameter $\phi_j^\star$. The above may alternatively be expressed as a convolution of the component-specific kernel $k(\cdot|\phi)$ with the discrete mixing measure $G^\star = \sum_{j=1}^{t} p_j^\star \delta_{\phi_j^\star}$:

$$f^\star(y) = \int k(y|\phi) G^\star(\mathrm{d}\phi). \tag{1.8}$$

Mixture models can be used for density estimation (Escobar and West (1995), Ferguson (1973)), regression (Müller et al. (1996)), and model-based clustering (Fraley and Raftery (2002)). When using mixture models for model-based clustering, one focuses on the groups to which each data point has been assigned, which naturally induce a partition in the data: two data points belong to the same cluster of the partition if and only if they have both been assigned to the same group. We denote this partition of the data by $A = (A_1, \ldots, A_{K_n})$, where K_n denotes the number of clusters in the partition. For a recent review on model-based clustering for mixture models, see Grün (2019). For a recent review on mixture models in general, see Fruhwirth-Schnatter et al. (2019).

In this thesis, we consider nonparametric mixture models, with nonparametric priors on the mixing measure G of Equation (1.8).

Consistency results for BNP mixtures

As described in Section 1.1.1 general consistency results are not necessarily applicable to nonparametric models. Extensive research has led to consistency in density estimation for Dirichlet process mixtures (Ghosal et al. (1999) Ghosal and Van Der Vaart (2007) Kruijer et al. (2010), and other types of priors (Lijoi et al. (2005)). Nguyen (2013) proves consistency for mixing measures for finite mixtures and for BNP mixtures, and provides their corresponding contraction rates.

It is important to realise that consistency of the posterior distribution for the data-generating density and even for the mixing measure does not imply consistency of the inferred number of clusters. Empirically, many researchers have observed that DP mixture posteriors tend to overestimate the number of clusters (West and Escobar, 1993; Lartillot and Philippe, 2004; Onogi et al., 2011). More recently, Miller and Harrison (2013, 2014) proved that the posterior distribution on the number of clusters does not concentrate to the number of components in DP and PYP mixtures. A possible explanation for this inconsistency result can be found in a result proved by Rousseau and Mengersen (2011), that in overfitted finite or infinite mixture models, the weights attributed to extra clusters go to zero as the number of observations grows. Provided that the weights for the extra components are infinitesimally small, any mixture can be approximated arbitrarily well by a mixture with a larger number of components.

Despite the above inconsistency results, it is possible to achieve posterior consistency for the number of clusters in DP and PYP mixtures. Guha et al. (2021) introduce a fast and simple post-processing procedure for DP mixtures which provides clustering consistency. Alamichel et al. (2022) extend this result to PYP mixtures. Ascolani et al. (2022) show that posterior consistency for the number of clusters can be achieved in certain cases for a DP mixture model by putting a prior on the DP concentration parameter α. DP mixtures modeled in this way can be considered as mixtures of DP mixtures (Antoniak, 1974) and are commonly used in practice.

Beyond the distribution over the number of clusters, an interesting question in cluster analysis is the distribution over the partition space across clusters induced by BNP mixture models. This space is large and complex: the number of possible clusterings of n items grows exponentially according to $B(n)$, the Bell number of n items (Bell (1934)). Since it would be infeasible to describe the posterior density of all the unique partitions, it is common practice to find a point estimator to concisely represent the posterior.

The optimal Bayes estimate of the clustering under the 0-1 loss function is equivalent to the maximum a-posteriori (MAP) clustering estimator (Binder (1978)), and is commonly used in Bayesian model-based procedures (Broët et al. (2002), Kim et al. (2006), Li et al. (2007)). The 0-1 loss function may be described intuitively as follows: no loss is incurred if the clustering estimate equals the true clustering and a loss of one is incurred for any other clustering estimate. Rajkowski (2019) investigate theoretical properties of the MAP partition in the particular case of Gaussian Dirichlet process mixture models (where the cluster means have Gaussian distribution and, for each cluster, the observations within the cluster have Gaussian distribution). Along with some nice theoretical properties, they prove that model mis-specification can lead to non-consistency of the MAP partition.

1.1.5 Bayesian computation

In a Bayesian analysis, many quantities of interest can be written in the form

$$I = \int h(\theta) d\Pi(\theta | y_{1:n}), \tag{1.9}$$

i.e. as an expectation of some quantity $h(\theta)$ with respect to the posterior distribution. For example, the expected value of θ under the posterior corresponds to the case where $h(\theta) = \theta$. Since posterior distributions are generally impossible to calculate directly, numerical methods must be used.

Even if the posterior distribution is unavailable in closed form, it can be possible to simulate from it. We can estimate I of Equation 1.9 using a Monte Carlo estimator $\hat{I}_T^{MC}$, defined as

$$\hat{I}_T^{MC} = \frac{1}{T}\sum_{t=1}^{T} h(\theta_t) \qquad \theta_t \sim_{iid} \Pi(\cdot|y_{1:n}).$$

Monte Carlo estimators have zero bias, and under the very general conditions of the Laws of Large Numbers, they converge to the truth, I, as the number of simulated data points T goes to infinity. Despite these nice properties, however, they are rarely used in practice: their variances can be quite large, and in many cases, it is impossible to simulate directly from the posterior distribution. An alternative solution is to use importance sampling, where data is simulated from an alternative sampling distribution, and weights are associated with each simulated data point to correct for the difference between the sampling distribution and the posterior distribution (see for example Geweke (1989)). Given a sampling distribution γ that emits a density (which we also denote γ) with respect to some measure, an importance sampling estimator $\hat{I}_T^{IS}$ is defined as

$$\hat{I}_T^{IS} = \frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta_t)\pi(\theta_t|y_{1:n})}{\gamma(\theta_t)}, \qquad \theta_t \sim_{iid} \gamma$$

The quality of $\hat{I}_t^{IS}$ depends crucially on the choice of γ, and is unbiased and consistent as long as the support of $\Pi(\cdot|\theta)$ is contained in the support of γ.

In the case of more complex models, it is often difficult to design an importance sampling distribution γ that places a large number of samples in regions of high posterior density. Classes of algorithms suitable for these situations include Markov chain Monte

Carlo (MCMC) algorithms and sequential Monte Carlo (SMC) algorithms.

The basic idea of MCMC methods is to simulate a Markov chain whose stationary distribution is the target posterior distribution. Monte Carlo estimates can then be formed using the elements in the chain (usually after discarding the first few elements of the chain, referred to as the "burn-in" time). Gibbs sampling (Gelfand and Smith (1990)) is an MCMC method useful when the parameter θ is of dimension greater or equal to two, and where it is possible to simulate from conditional distributions of its components. Each transition of the Markov chain involves resampling a component of the parameter vector from its conditional distribution.

The Metropolis Hastings method (Robert et al. (1999)) first introduced by Metropolis and Ulam (1949) is an MCMC method based on using a suitable transition kernel and acceptance probability in order to ensure a "detailed balance" condition, which is necessary for the stationary distribution of the Markov chain to be equal to the target distribution. A large number of extensions have been introduced, for example, adaptive versions (Roberts and Rosenthal (2009)), where the best choice of the parameters is learned during the convergence of the chain, and parallel tempering (Geyer (1991)) which uses a sequence of Markov chains running in parallel with earlier chains in the sequence easier to sample from, with neighbouring chains close in distribution, and which allows neighbouring chains to swap state in order to improve mixing. The last chain of the sequence has stationary distribution equal to the target.

SMC methods also involve the construction of a sequence of intermediate distributions $f_0(\theta), f_1(\theta), \dots, f_T(\theta)$, in such a way that the final distribution $f_T(\theta)$ corresponds to the Bayesian posterior target distribution $\Pi(\cdot|y_{1:n})$. In SMC, one simulates M particles on Θ which are then propagated from f_0 to f_T so that in the end, one obtains sets of vectors of the form $(\theta_0^{(m)}, \theta_1^{(m)}, \dots, \theta_T^{(m)})$ for all $m \in \{1, \dots, M\}$, where $\theta_i^{(m)}$ is distributed according to f_i for any i in $\{0, \dots, T\}$ and for any $m \in \{1, \dots, M\}$. In particular, the set $\{\theta_T^{(1)}, \dots, \theta_T^{(M)}\}$ will be a sample from the target posterior distribution, from which Monte Carlo estimates may be made.

For $t \in \{1, \dots, T\}$ a common choice for the distribution f_t is the partial posterior distribution $\Pi(\cdot|y_{1:t})$, and a common choice for f_0 is the prior distribution $\Pi(\cdot)$. At time step t, each of the particles can then be propagated from f_{t-1} to f_t using importance sampling, with sampling distribution $\Pi(\cdot|y_{1:t-1})$ and with weight $p(y_t|\theta)$, in such a way that the re-weighted sample will be distributed according to $\Pi(\cdot|y_{1:t})$. Practically, an additional re-weighting step is necessary to avoid particle degeneracy (i.e. to avoid the weights of all but one particle becoming so small that they no longer have an adequate influence on the final MC sample. It has been proved that SMC is guaranteed to fail without this step, see Liu et al. (1998) and Liu and Liu (2001)). Pseudocode for a very basic SMC sampler, the Bootstrap sampler, is provided in Algorithm 1.

Due to the increasing availability and power of computational resources, SMC methods have become very popular since the 1990s, and there is a rich literature on the construction of their algorithms (see for example Gilks and Berzuini (2001), Neal (2001), Doucet et al. (2001), and Chopin (2002), among others). Unlike other alternatives such as MCMC algorithms, SMC algorithms can process the data sequentially. This makes them perfectly suited to situations where the dimension of the data is large and it would be unrealistic to process all of it in one go, or where data arrives sequentially and one wishes to make estimates online.

Algorithm1: Bootstrap Filter

1. initialisation, $t = 0$

   - For $i = 1, \ldots, M$ sample $\theta_0^i \sim \Pi(\cdot)$ and set $t = 1$.

2. Importance sampling step

   - For $i = 1, \ldots, M$, sample $\tilde{\theta}_t^i \sim \Pi\left(\cdot | y_{1:(t-1)}\right)$ and set $\tilde{\theta}_{0:t}^i = \left(\theta_{0:t-1}^i, \tilde{\theta}_t^i\right)$.
   - For $i = 1, \ldots, M$ evaluate the importance weights

$$\tilde{w}_t^i = \frac{p(y_t|\tilde{\theta}_t^i)}{\sum_{i=1}^{M} p(y_t|\tilde{\theta}_t^i)}.$$

3. Selection step

   - Resample with replacement M particles $(\theta_{0:t}^i; i = 1, \ldots, M)$ from the set $\left(\tilde{\theta}_{0:t}^i; i = 1, \ldots, M\right)$ according to the importance weights.
   - Set $t \leftarrow t + 1$ and go to step 2.

1.1.6 Approximate Bayesian computation (ABC)

In the case of highly complex or high-dimensional problems, the simulation-based methods described above can be too expensive computationally. In certain situations, they can even be mathematically impossible. For example, when the likelihood function cannot be evaluated, one has no way of calculating acceptance probabilities for the Metropolis-Hastings algorithm, or of calculating weights of the SMC algorithm. Furthermore, the decomposition required for Gibbs sampling is generally unavailable. In these cases, a more approximate method of inference is necessary. Previous solutions to these problems include Laplace approximations (Tierney and Kadane (1986)) and variational Bayes methods (Jaakkola and Jordan (2000)). However, Laplace approximations require unrealistic prior knowledge of the posterior distribution, and variational Bayes models replace the true posterior distribution with a much simpler pseudo-posterior which may fail to capture important features of the true model.

Approximate Bayesian computation (ABC) methods, proposed by Tavaré et al. (1997) in the context of population genetics, have been referred to by Marin et al. (2012) as "the most satisfactory approach to intractable likelihood problems." The basic idea is simple. Instead of evaluating the likelihood function, one simulates a set of M parameter and data pairs from the prior and likelihood function $\left((\theta^1, z^1), \ldots, (\theta^M, z^M)\right)$. For all of the $i's$ for $i \in \{1, \ldots, M\}$ for which z^i is sufficiently "close" to the observed data y, the parameter θ^i is stored. The set of stored parameters is a Monte Carlo sample from the ABC pseudo-posterior distribution. In the special case where by "close" we mean "equal to", the pseudo-posterior distribution will be the true posterior distribution. Otherwise, it will be some approximation of the true posterior distribution. The most basic accept/reject ABC algorithm (Tavaré et al. (1997)) given in Algorithm 2.

Algorithm2: Accept/reject ABC

1. Simulate θ^i $(i = 1, \ldots, M)$ from $\Pi(\cdot)$
2. Simulate $z^i = (z^i_1, \ldots, z^i_n)$ $(i = 1, \ldots, M)$ from $P(\cdot|\theta^i)$
3. Accept the θ^i satisfying $|\eta(y^i) - \eta(z^i)| \leq \epsilon$, where $\eta(\cdot)$ is a statistic and ϵ is a tolerance level.

Defining an approximation to the likelihood as

$$\tilde{p}_{\epsilon,\theta}\left(\eta(y)\right) := \int \mathbf{1}_{\{\|\eta(y)-\eta(z)\| \leq \epsilon\}} dP(z|\theta), \tag{1.10}$$

this ABC accept/reject algorithm produces samples from the following pseudo-posterior density (when marginalizing out the simulated data $\eta(z)$):

$$\pi_\epsilon(\theta) \propto \pi(\theta)\tilde{p}_{\epsilon,\theta}\left(\eta(y)\right). \tag{1.11}$$

ABC methods are relatively recent and have enjoyed a large amount of research attention in the past two decades, both from a theoretical point of view and from a computational point of view.

As with any approximate statistical method, it is crucial to understand the asymptotic behavior of ABC posterior estimations in order to validate their reliablilty. Li and Fearnhead (2018b) and Frazier et al. (2018) have considered the asymptotic properties of ABC, with ABC tolerances decreasing as the number of observations goes to infinity. Both papers have shown that convergence of ABC point estimators depends on the relationship between the rate of convergence of the summary statistics and that of the tolerance.

Frazier et al. (2018) have additionally proved posterior consistency of the ABC posterior, and both Frazier et al. (2018) and Li and Fearnhead (2018a) have proved results on the asymptotic shape of the ABC posterior distribution. These results again depend on the relationship between the rate of convergence of the summary statistics and that of the tolerance. In particular, posterior consistency is only proved in the case where all summary statistics converge at a rate that is much faster than that of the tolerance. The shape of the asymptotic ABC posterior distribution is only proved in situations where all dimensions of the summary statistics converge at the same rate as each other: either they all converge at a "slow" rate relative to ϵ, or they all converge at a "fast" rate relative to ϵ.

1.2 Contributions and thesis outline

This thesis comprises three main chapters, each representing independent work. Chapter 2 focuses on the asymptotic properties of ABC methods (and is based on work in preparation for submission to the Electronic Journal of Statistics). Chapter 3 considers the natural language processing application of Bayesian nonparametric models (and is unsubmitted, unpublished work). Chapter 4 presents consistency results for Bayesian nonparametric mixture models (and is partially based on work published at the Advances in Approximate Bayesian Inference conference and partially based on work in progress). In the next subsections, I briefly summarise the methodologies, the results, and the impact of each of the following chapters. I also outline which parts were carried out as part of a collaboration.

1.2.1 Asymptotic properties of ABC

In Chapter 2, I present joint work with C. Robert, J. Rousseau and R. Ryder on the asymptotic properties of ABC methods. We consider the ABC posterior as the tolerance parameter ϵ decreases with increasing data: as ϵ converges to zero, the ABC summary statistics converge to their limiting values. As outlined in Subsection 1.1.6, previous results have focused on settings where the summary statistics all converge at the same rate. Either they all converge "slowly" relative to ϵ, or they all converge "quickly" relative to ϵ. We generalise to the more realistic case where different components of the summary statistics converge at different rates. We prove posterior consistency under this set-up, even when certain summary statistics do not converge at all, and provide a closed-form expression for the asymptotic ABC posterior distribution. We prove similar but stronger results in the case of ABC after post-processing.

I was the lead researcher for the two theorems and the simulation study that deal with ABC before post-processing. This was the first project of my DPhil, and I had guidance from C. Robert, J. Rousseau and R. Ryder. The results regarding ABC posteriors after post-processing are due to J. Rousseau and R. Ryder, and are not my work.

1.2.2 Bayesian nonparametric models for grammars

In Chapter 3, I present joint work with R. Ryder and J. Rousseau on Bayesian modelling of grammars. We focus on the class of 2-multiple context-free grammars (2-MCFGs). As discussed in Subsection 1.1.3, although the majority of work on grammar modelling in the literature has focused on the class of context-free grammars, more general classes would be necessary in order to capture all features of natural language. 2-MCFGs allow, for example, cross-dependencies across words in sentences, a feature which is impossible to capture using context-free grammars. We propose a model for 2-MCFGs based on the hierarchical Dirichlet process, inspired by previously proposed models for context-free grammars. We develop a sequential Monte Carlo algorithm to make inferences and illustrate our method with synthetic and real data.

I was the lead researcher for this project, both for modelling and implementation. Many ideas came from regular discussions with R. Ryder and J. Rousseau.

1.2.3 Consistency results for BNP mixtures

In Chapter 4, I present joint work with J. Arbel, L. Alamichel and G. Kon Kam King on the topic of consistency of Bayesian nonparametric mixture models. First, we consider the question of consistency for the number of clusters. In our first theorem, we prove that, unlike the case of DP mixture models, PYP mixture models (which generalise DP mixture models) remain inconsistent for the number of clusters when a prior is put on α, in the special case where the true number of components in the data generating mechanism is equal to 1 and the discount parameter $\sigma \in (0, 1)$ is a fixed constant. We illustrate this result with a simulation study. Secondly, we consider the MAP estimator over the space of partitions. We prove consistency of the MAP partition for DP mixture models when the concentration parameter, α_n goes deterministically to zero, and when the true partition is made of only one cluster. This second result is complimentary to that of Rajkowski (2019) who prove a negative result in the case of a fixed concentration parameter.

I was the overall lead researcher for this project. The first theorem was joint work with J. Arbel, and the second theorem was my individual work. The simulation study included in the Appendix section was carried out by L. Alamichel and G. Kon Kam King and was not my work.

Chapter 2

Asymptotic properties of approximate Bayesian computation (ABC)

Abstract

Approximate Bayesian computation (ABC) is a computational method suitable for computing posterior distributions in situations highly complex models, where the likelihood function is intractable but can be simulated from. Previous authors have proved consistency and provided rates of convergence in the case where all summary statistics converge at the same rate as each other. We generalize to the case where summary statistics may converge at different rates, and provide an explicit representation of the shape of the ABC posterior distribution in our general setting. We also show under our general setting that local linear post-processing can lead to significantly faster contraction rates of the pseudo-posterior.

2.1 Introduction

Likelihood-free methods in Bayesian statistics are methods for posterior inference in situations where likelihoods are intractable or unavailable in closed form, but may be simulated from. Such situations are typical of real-life applications, when models are defined by complex generative processes. This can result in likelihoods involving high dimensional integrals which are impossible to compute in a reasonable amount of time. Examples of such likelihood-free methods include simulated methods of moments (Duffie and Singleton, 1990), indirect inference (Gourieroux et al., 1993), synthetic likelihood (Wood, 2010) and approximate Bayesian computation (ABC) (Sisson et al., 2018). In this work, we focus on the latter. At its core, ABC relies on simulating many data sets from the prior predictive. The data sets are summarized by a low dimensional statistic, and only those within a small pseudo distance (the tolerance) of the observed data are kept.

ABC was first introduced in the context of population genetics (Pritchard et al., 1999). It has since then been applied in research areas as diverse as population genetics (Pritchard et al., 1999), protein networks (Ratmann et al., 2007), epidemiology (Tanaka et al., 2006) , inference for extremes (Bortot et al., 2007), dynamical systems (Toni et al., 2009), and Gibbs random fields (Grelaud et al., 2009). Due to its increasing popularity in applied statistics, recent research has focused on the theoretical properties of ABC methods.

Fearnhead and Prangle (2012) consider the question of summary statistic choice, and find that summary statistics should ideally have the same dimension as the parameter to be estimated. Li and Fearnhead (2018b) and Frazier et al. (2018) have considered the asymptotic properties of ABC, with ABC tolerances decreasing as the amount of information in the data goes to infinity. Both papers have shown that convergence of ABC posteriors depends on the relationship between the rate of convergence of the summary statistics and that of the tolerance and they have proved results on the asymptotic shape of the ABC posterior distribution. These results again depend on the relationship between the rate of convergence of the summary statistics and that of the tolerance. In particular, posterior consistency is only proved in the case where all summary statistics converge at a rate that is much faster than that of the tolerance. The shape of the asymptotic ABC posterior distribution is only proved in situations where all dimensions of the summary statistics converge at the same rate.

In this work, we extend the results of Frazier et al. (2018) to the case where different components of the summary statistics converge at different rates, with some possibly not converging at all. We first prove consistency of the ABC posterior where different components of the summary statistics are allowed to converge at heterogeneous rates. We next prove a general result on the asymptotic shape of the ABC posterior in the same context and our results cover the more realistic case where certain summary statistics do not converge at all.

A well known technique to reduce the curse of the dimension of the set of summary satistics is based on non linear regressions, typically a post processing step, as introduced by Blum and François (2009); see also Blum (2010). Recently Li and Fearnhead (2018a) have shown, in the special case of asymptotically normal summary statistics all concentrating at the same rate, that the local linear postprocessing step proposed in Blum and François (2009) leads to a significant improvement in the behaviour of the ABC posterior. However since the post - processing step aimed at reducing the impact of the dimension of the summary statistics, it is important to study its efficiency in a context where the summary statistics are not as well behaved as considered in Li and Fearnhead (2018a). We fill this gap by showing that local linear post-processing induces significant improvement even when summary statistics have heterogeneous behaviour.

In Section 2.2 we provide details of our set-up, state the assumptions that we will be using and introduce key notation. In Section 2.3 we state our result on the asymptotic form of the ABC posterior and in Section 2.4 we study its consequence on the local linear regression post-processing strategy. In Section 2.5 we illustrate these results empirically. A short discussion is provided in Section 2.6. The details of proofs of theoretical results are left to the appendix.

2.2 Background

We observe data $y \in \mathbb{R}^n$ and assume that they arise from the model $\{P_\theta(\cdot) : \theta \in \mathbb{R}^d\}$, where $P_\theta(\cdot)$ is a density function. We denote by $\theta_0 \in \mathbb{R}^d$ the unknown true value of interest that generated the observed data y. We denote by $\pi(\cdot)$ the prior density on parameter space, and by $\Pi(\cdot)$ the corresponding cumulative density function.

The idea of ABC is to make inference on the posterior distribution using Monte Carlo samples of parameter-data pairs $(\theta_i, z_i) \in \mathbb{R}^d \times \mathbb{R}^n$, simulated from the forward model. Distances between observed data y and simulated data z_i determine the role θ_i will play in the estimation of θ_0.

When the dimension of data n is large, it is inefficient to compute distances on the raw data. It is thus common practice to instead compute distances between lower dimension summary statistics of the observed and simulated data. We thus define a summary function $\eta : \mathbb{R}^n \to \mathbb{R}^k$ from data space to summary space, where $k < n$. Summary statistics will typically be sample moments or quantiles of the data, although many other summary statistics have been considered in the literature. In this work, we consider the Euclidean distance.

Although more sophisticated ABC algorithms now exist, we will focus on the simple accept/reject ABC algorithm (Tavaré et al., 1997) ; Pritchard et al. (1999), described in Algorithm 1 below. Algorithm 1 generates a *reference table* from the model consisting of (parameter, summary statistic) pairs. For a given tolerance level ϵ, parameters corresponding to data within a distance ϵ of the observed data are accepted. Parameters corresponding to data further away from ϵ of the observed data are rejected. The accepted values form a sample of the ABC posterior distribution which is then used to estimate quantities of interest by Monte Carlo.

Frazier et al. (2018) and Li and Fearnhead (2018b) suggest that the tolerance, ϵ should depend on n, the dimension of the data, and should tend to zero as n goes to infinity. Indeed, as n increases, information about underlying parameters accumulates in samples. If a simulated parameter is close in distance to θ_0, than data generated from

```
Observed data $y$, summary statistics $\eta$, threshold $\epsilon$ for $i = 1, \ldots, M$ do
end
Simulate $\theta^i \sim \Pi(\cdot)$ Simulate $z^i = (z^i_1, \ldots, z^i_n) \sim P_{\theta^i}(\cdot)$ if $\|\eta(y) - \eta(z^i)\| \leq \epsilon$
 then
end
Accept $\theta_i$
```

Algorithm 1: Accept/reject ABC

it should be close in distance to the observed data, y. Hereafter, we will thus denote the ABC tolerance by ϵ_n, and let $\lim_{n\to\infty} \epsilon_n = 0$.

Defining an approximation to the likelihood as

$$\tilde{p}_{\epsilon_n,\theta}\left(\eta(y)\right) := \int \mathbf{1}_{\{\|\eta(y)-\eta(z)\|\leq\epsilon_n\}} dP_\theta(z), \tag{2.1}$$

this ABC accept/reject algorithm produces samples from the following pseudo-posterior distribution (when marginalizing out the simulated data $\eta(z)$):

$$\pi_{\epsilon_n}(\theta) \propto \pi(\theta)\tilde{p}_{\epsilon_n,\theta}\left(\eta(y)\right). \tag{2.2}$$

We will show that properties of the asymptotic ABC posterior depend on the relationship between ϵ_n and the rate at which summary statistics converge to some well-defined limit. We formalize the notion of convergence of summary statistics in Assumption 1 below.

Assumption 1. *There exists some Lipschitz continuous mapping $b : \mathbb{R}^d \to \mathbb{R}^k$ such that, for all $1 \leq j \leq k$, there exists some sequence v_{nj} such that*

$$v_{nj}\left(b(\theta)_j - \eta(z)_j\right) = O_P(1).$$

Without loss of generality, we assume $v_{n1} \leq v_{n2} \leq \ldots \leq v_{nk}$.

In this work, we consider the novel setting where, for some $1 \leq k_0 < k$, the statistics from 1 to k_0 converge at slow rates and the statistics from $k_0 + 1$ to k converge at fast rates, i.e.

$$\lim_{n\to\infty} v_{nj}\epsilon_n = 0 \qquad \forall 1 \leq j \leq k_0$$
$$\lim_{n\to\infty} v_{nj}\epsilon_n = \infty \qquad \forall k_0 < j \leq k.$$

Throughout this chapter, for any vector g, we let $g_{(1)} = (g_1, \ldots, g_{k_0})$ denote the vector of length k_0 of all components g_j of g with $j \leq k_0$. We let $g_{(2)} = (g_{(k_0+1)}, \ldots, g_k)$ denote the vector of length $(k - k_0)$ of all components g_j of g with $j \geq (k_0 + 1)$. We also write $D_n = \text{diag}(v_{n,j}, j \leq k)$; $D_{n,1}$ its upper k_0 sub-matrix and $D_{n,2}$ its lower $k - k_0$ sub-matrix. Set $Z_n(\theta)$ the random variable $D_n(\eta(z) - b(\theta))$ with $z \sim P_\theta$.

In addition to Assumption 1, our results rely on the following assumptions, which we discuss in Remarks 1 and 2. We show that these assumptions are verified on a toy example in Example 1.

Assumption 2. *The matrix $\nabla b_{(2)}(\theta_0)$ is of full rank, where ∇ represents the gradient operator.*

Assumption 3. *There exist some constant $\delta > 0$, some strictly positive bounded Lipschitz continuous function $\gamma : \mathbb{R}^{k_0} \to \mathbb{R}^+$, some constant $R > 0$, and some $o(1)$ sequence L_n such that, for all compact sets $K \in \mathbb{R}^{k_0}$, and for all $1 \geq t > 0$,*

$$\sup_{\|\theta-\theta_0\|<\delta} \sup_{m\in K} \left| \frac{P_\theta\left(\sum_{j=1}^{k_0} \frac{(Z_{n,j}(\theta)-m)^2}{v_{n,j}^2} < t^2\epsilon_n^2\right)}{t^R L_n} - \gamma(m) \right| = o(1).$$

Assumption 4. *There exist $\kappa > d$, some function $c : \Theta \to \mathbb{R}^+$, such that the following is true for all $n \geq 1$, for all $u > u_0$ and for all $\theta \in \Theta$.*

$$\max_{j\geq k_0+1} P_\theta\left(|Z_{n,j}(\theta)| > u\right) \leq c(\theta)u^{-\kappa}; \qquad \int_\Theta c(\theta)\pi(\theta)d\theta < \infty$$

Assumption 5. *For all $M > 0$, there exist monotone nonincreasing functions $M_2, > 0$, $\kappa > d$ and a non decreasing sequence $u_n \geq u_0 > 0$ such that, $u_n = o(v_{n,k_0+1}\epsilon_n)$ and for all n large enough, $u \geq u_n$, $\|\theta - \theta_0\| \leq \delta_n$, and $j > k_0$*

$$\sup_{\|z\| \leq M} P_\theta \left(|Z_{n,j}(\theta)| > u | Z_{n,(1)}(\theta) = z \right) \leq M_2 u^{-\kappa},$$

with

$$\left(\frac{v_{n,k_0}}{v_{n,k_0+1}} \right)^{\kappa - d} (\epsilon_n v_{n,k_0+1})^{-d} = o(L_n)$$

3.

Remark 1. *Assumption 2 guarantees that enough summary statistics have a fast rate of convergence. Assumption 3 guarantees sufficient stability of the summary statistics with slow rates of convergence. Typically, L_n will be a sequence proportional to $\prod_{i=1}^{k_0} v_{ni}\epsilon_n$ and R will equal to k_0. (See Example 1 below). Assumption 4 controls the tail behavior of the summary statistics with fast rates of convergence. Assumption 4 is similar to the tail assumption Frazier et al. (2018) and we refer to their discussion on this assumption. It holds true in particular if $\limsup_n E_\theta(\|Z_{n,(1)}\|^\kappa) < \infty$. Assumption 5 is implied by Assumption 3 and Assumption 4 in the particular case where the vectors of slow and fast converging summary statistics are mutually independent, but holds more generally than that.*

Remark 2. *Our Assumptions are slightly weaker than those of Li and Fearnhead (2018b) and of Frazier et al. (2018). While in Assumption 1 we require just stochastic boundedness, these authors require central limit theorems for the summary statistics. More importantly we do not impose, as in Li and Fearnhead (2018b) that the summary statistics concentrate at the same rate nor do we impose some limiting distribution for the fast ones. However γ in Assumption 3 can be thought of as the limiting density of the the slow (first k_0) summary statistics. Compared to Frazier et al. (2018), we do not impose that ϵ_n is either smaller or larger than all summary statistics. While our Assumption 3 and Assumption 4 apply only to the summary statistics with slow and* *fast rates of convergence respectively, Frazier et al. (2018) make similar assumptions for the full vector of summary statistics.*

Example 1. *We verify all of the assumptions for a simple uniform toy example. We assume that the data follows a continuous unit uniform distribution with unknown location parameter $\theta \in \mathbb{R} : z_i \sim U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right), i \in \{1, \dots, n\}$. We put a standard uniform prior on the parameter: $\theta \sim U(0,1)$. We take the first k_1 summary statistics to be the first k_1 observations. These do not converge, so we have $b(\theta)_i = 0$ and $v_{ni} = 1$ for $i \in \{1, \dots, k_1\}$. We take $\eta_{k_1+1}(z) = \bar{z}_n = \sum_{i=1}^n z_i/n$ and $\eta_{k_1+2}(z) = \max_{i \leq n} z_i$. The statistic $\eta(z)_{k_1+1}$ converges at the rate $v_{n,k_1+1} = \sqrt{n}$ to θ and $\eta(z)_{k_1+2}$ at the rate $v_{n,k_1+2} = n$ to $1/2 + \theta$. We consider $\epsilon_n = o(1)$ and we consider two scenarii. Scenario 1: $1/n << \epsilon_n << 1/\sqrt{n}$ Scenario 2: $1/\sqrt{n} << \epsilon_n << o(1)$.*

In Scenario 1 $k_0 = k_1 + 1$ and $\eta_{(1)}(z) = (z_1, \dots, z_{k_1}, \bar{z}_n) \in \mathbb{R}^{k_1+1}$ while in scenario 2, $k_0 = k_1$ and $\eta_{(1)}(z) = (z_1, \dots, z_{k_1}) \in \mathbb{R}^{k_1}$. Assumptions 1 and 2 are trivially verified in both scenarii. Using the same notation as before, we are then in the setting where $\forall i \leq k_0, \lim_{n\to\infty} v_{ni}\epsilon_n = 0$ and $\lim_{n\to\infty} v_{n(k_0+1)}\epsilon_n = \infty$.

We now prove that assumptions 3-5 are verified. We treat the case of scenario 1, which is more difficult.

Let K_1 be a compact subset of $(-1/2, 1/2)^{k_1}$ and $K = K_1 \times [-M, M]$. We have $Z_{n,(1)} = (z_1 - \theta, \cdots, z_{k_1} - \theta, \sqrt{n}(\bar{z}_n - \theta))$ and $\bar{Z}_{n,(2)} = n(\eta_{k_1+2} - \theta - 1/2)$. Let $m \in K$

$$P_\theta\left(\left\|(Z_{n,(1)}(\theta) - m\right\| \leq t\epsilon_n\right) = E_\theta\left(P_\theta((\bar{z}_n - m_{k_1+1})^2 \leq t^2\epsilon_n^2 - \sum_{j=1}^{k_1}(Z_{n,j} - m_j)^2 | z_1, \cdots, z_{k_1}\right)$$

$$= 2\left(E_\theta\left(\mathbb{1}_{\sum_{j=1}^{k_1}(Z_{n,j}-m_j)^2 \leq t^2\epsilon_n^2}\sqrt{t^2\epsilon_n^2 - \sum_{j=1}^{k_1}(Z_{n,j} - m_j)^2}\varphi(m_{k_1+1})(1 + o(1))\right)\right)$$

$$= 2\varphi(m_{k_1+1})\prod_{i=1}^{k_1}(f_\theta(m_i) + o(1))\int \mathbb{1}_{\sum_{j=1}^{k_1}(z_j-m_j)^2 \leq t^2\epsilon_n^2}\sqrt{t^2\epsilon_n^2 - \sum_{j=1}^{k_1}(z_j - m_j)^2 dz_1 \cdots dz_{k_1}}$$

$$= \varphi(m_{k_1+1})\prod_{i=1}^{k_1}(f_\theta(m_i) + o(1))(t\epsilon_n)^{k_1}\int_0^1 u_1^k(1 - u^2)du$$

and Assumption 3 is verified, with $L_n = 2\epsilon_n^{k_1} \int_0^1 u_1^k(1-u^2)du$, *and* $\gamma(m) = \varphi(m_{k_1+1}) \prod_{i=1}^{k_1} f_\theta(m_i)$.

We also have

$$P_\theta\left(|n(\eta_{k_1+2} - \theta - 1/2)| \geq u\right) = Pr\left(\max_i (z_i - \theta) \leq 1/2 - u/n\right) = (1-u/n)^n \leq e^{-u}$$

so that Assumption 4 holds for all $\kappa > 0$.

Here $Z_{n,(2)}$ *and* $Z_{n,(1)}$ *are not independent, but we can verify Assumption 5. First note that, with* $z_{(n)} = \max_i z_i$ *and for all* $m \in \mathbb{R}^{k_1+1}$, *writing*

$$\begin{aligned} P_\theta(n|z_{(n)} - \theta - 1/2| > u | Z_{n,(1)} = z) &= P_\theta(z_{(n)} \leq \theta + 1/2 - u/n | Z_{n,(1)} = m) \\ &\leq P_\theta(\max_{i \geq k_1+1} z_i \leq \theta + 1/2 - u/n | Z_{n,(1)} = m) \\ &= P_\theta(\max_{i \geq k_1+1} z_i \leq \theta + 1/2 - u/n | \bar{z}_n - \theta = m_{k_1+1}/\sqrt{n}) \end{aligned}$$

Hence without loss of generality we can work with $k_1 = 0$. *Consider the change of variables* $x_i = z_i - \theta$ *for* $i \geq 2$ *and* $x_1 = \bar{z}_n - \theta$, *whose distribution is independent of* θ. *The joint density of* $x = (x_1, \cdots, x_n)$ *if given by*

$$f_x(x) = \prod_{i \geq 2} \mathbb{1}_{(-1/2,1/2)}(x_i) \mathbb{1}_{nx_1 - (n-1)\bar{x}_{n-1} \in (-1/2,1/2)}, \quad \bar{x}_{n-1} = \sum_{i \geq 2} x_i/(n-1)$$

This leads to for all $m \in \mathbb{R}$,

$$P\left(\max_{i \geq 2} x_i \leq 1/2 - u/n | x_1 = m/\sqrt{n}\right) = \frac{\int_{[-1/2,1/2-u/n]^{n-1}} \mathbb{1}_{\sqrt{n}m - (n-1)\bar{x}_{n-1} \in (-1/2,1/2)} dx_2 \cdots dx_n}{\int_{[-1/2,1/2]^{n-1}} \mathbb{1}_{\sqrt{n}m - (n-1)\bar{x}_{n-1} \in (-1/2,1/2)} dx_2 \cdots dx_n} \tag{2.3}$$

We also have that

$$\begin{aligned} &\int_{[-1/2,1/2]^{n-1}} \mathbb{1}_{\sqrt{n}m - (n-1)\bar{x}_{n-1} \in (-1/2,1/2)} dx_2 \cdots dx_n \\ &= P(|\sqrt{n-1}\bar{x}_{n-1} - m\sqrt{n/(n-1)}| \leq 1/(2\sqrt{n-1})) \gtrsim \frac{e^{-6m^2}}{\sqrt{n}}, \end{aligned}$$

which plugged into (2.3) *implies*

$$P\left(\max_{i\geq 2} x_i \leq 1/2 - u/n | x_1 = m/\sqrt{n}\right) \lesssim \sqrt{n} e^{6m^2} \int_{[-1/2, 1/2-u/n]^{n-1}} dx_2 \cdots dx_n$$
$$\lesssim \sqrt{n} e^{6m^2} e^{-u} \lesssim e^{6m^2} e^{-u/2} \quad \forall u \geq \log n := u_n$$

Therefore as soon as $\epsilon_n >> \log n/n$, *Assumption* 5 *holds true for all* $\kappa > 0$.

2.3 Main theorems

Our first result, Theorem 2.3.0.1 below is a Bayesian consistency result. It asserts that the ABC posterior density of any set which does not include the parameter which generated the observations, θ_0, behaves like an $o_P(1)$ random variable. Since the ABC posterior will differ from the true posterior given the observations, such a result is crucial if one wishes to quantify uncertainty based on the ABC posterior.

Theorem 2.3.0.1. *Under Assumptions 1, 2, 3, and 4 our ABC posterior distribution concentrates: There exists some monotone decreasing sequence* λ_n *with* $\lambda_n \to 0$ *such that*

$$\int_{\|\theta-\theta_0\|\geq\lambda_n} \pi_{\epsilon_n}(\theta|\eta(y)) d\theta = o_P(1).$$

The rate of concentration of the ABC posterior, λ_n, is of the same order as the sequence $\bar{\lambda}_n$ of Assumption 4. Thus, following Remark 1 on the form of the function $\rho(\cdot,\cdot)$ and the sequence L_n, we typically will have that, the faster the rate of the convergence of the fast statistics, $v_{n(k_0+1)}$, the faster the rate λ_n will be. The greater the quantity of slow converging statistics, k_0, and the slower the slow converging statistics converge, the slower the rate λ_n will be.

Our second result, Theorem 2.3.0.2 completely characterises the shape of the ABC posterior.

Theorem 2.3.0.2. *Under Assumptions 1, 2, 3, 4 and 5, the asymptotic ABC posterior may be expressed in closed form as*

$$\pi_{\epsilon_n}(\theta|\eta(y)) \propto \mathbb{1}_{\left\{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\| \leq \epsilon_n\right\}} \left(1 - \frac{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\|^2}{\epsilon_n^2}\right)^{\frac{R}{2}}, \qquad (2.4)$$

where R is the positive constant defined in Assumption 3.

As discussed in Remark 1, R will typically be equal to k_0, the number of summary statistics which converge at the slow rate. In the special case where $R = 0$, the shape of the ABC posterior distribution simplifies to a uniform distribution over the ellipsoid $\{\theta : \left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\| \leq \epsilon_n\}$. This is consistent with results in Frazier et al. (2018) and in Li and Fearnhead (2018a) for the case where all statistics converge at the fast rate (i.e. where $k_0 = 0$).

Redundant summary statistics which do not converge at all play exactly the same role on the shape of the asymptotic ABC posterior as summary statistics which converge at the slow rate.

The larger R is, the more concentrated the theoretical mass of (2.4) will be around θ_0. However, we will see in Lemma 2.7.1.3 that large R leads to low acceptance rate in Algorithm 1, and thus high Monte Carlo error.

Interestingly, the number of summary statistics which converge at the fast rate (i.e. $k - k_0$) will have no impact on the the rate of posterior concentration nor on the the shape of the ABC posterior (beyond the requirement $(k - k_0) > d$ by Assumption 2).

2.4 Local linear regression correction

ABC practitioners routinely use post-processing to improve the quality of the pseudo-posterior. Beaumont2002 introduced the idea of a local linear regression on the ABC output; empirical studies have since shown that this post-processing step can vastly ameliorate the pseudo-posterior, for a negligible computational overhead. In this section, we give results on the asymptotic behaviour of the regression-adjusted pseudo-posterior.

In general, post-processing corrections use the following idea: Consider the *pseudo model*

$$\theta = m(S) + u, \quad (\theta, S) \sim \pi_{\epsilon_n}(\theta, dS) \propto \pi(d\theta)P_\theta(dS)\mathbb{1}_{|S-S_0|\leq\epsilon_n}$$

and samples (θ_t, S_t) from the above distribution. Our distribution of interest is the distribution of $m(S_0) + u$. To approximate it, we learn the model m and consider the residuals $\hat{u}_t = \theta_t - m(S_t)$; the corrected ABC posterior samples are:

$$\theta_t = \theta_t + m(S_0) - m(S_t).$$

The regression adjustment we consider here corresponds to the locally linear model case, where $m(S) = (B,\, \beta_0)^T(S,\, 1)$ so that the targeted $B \in \mathbb{R}^{d\times k}, \beta_0 \in \mathbb{R}^d$ minimizes

$$L(B, \beta_0) = E_n[\|\theta - \theta_0 - \beta_0 - B^T(S - S_0)\|^2] \tag{2.5}$$

Let $\beta(j)$ the jth row of B, $\beta_{(1)}(j)$ (resp. $\beta_{(2)}(j)$) the first k_0 components of $\beta(j)$ (resp. the last $k - k_0$) .

In addition to the assumptions 1-5 we also assume the following:

A1 The function $\theta \to E_\theta(Z_n(\theta))$ is locally Lipschitz on a neighbourhood of θ_0 with Lipschitz constant L.

A2 There exists $\epsilon > 0$ such that

$$\sup_{\|\theta-\theta_0\|\leq\epsilon} E_\theta\left(\|Z_n(\theta)\|^4\right) < \infty.$$

We then have the following theorem.

Theorem 2.4.0.1. *Under assumptions 1-5 and assuming that [A1] and [A2] above hold, then on a set of y whose probability goes to 1, any minimizer in B of $L(B, \beta_0)$*

verifies

$$C_{11} D_{n,1}^{-1} \beta_{(1)}(j) = O(\epsilon_n / v_{n,1}), \quad \nabla b_2(\theta_0)^T \beta_{(2)}(j) = e_j + O(1/v_{n,1}), \tag{2.6}$$

where e_j *is the* j*-th vector in the canonical basis of* $\mathbb{R}^d$ *and*

$$C_{1,1} = E_n((Z_{n,(1)}(\theta) - E_n(Z_{n,(1)}(\theta)))(Z_{n,(1)}(\theta) - E_n(Z_{n,(1)}(\theta)))^T).$$

Let B^* *be the limit of* $B(j)$ *with minimal* L_2 *norm (by rows), then* $B^*(j) = (0, \cdots, 0, \Gamma_2 e_j)$ *with* $\Gamma_2 = \nabla b_2(\theta_0)[\nabla b_2(\theta_0)^T \nabla b_2(\theta_0]^{-1}$.

Moreover if

$$\theta' = \theta - B^*(S - S_0), \quad with(\theta, S) \sim \pi_{\epsilon_n}(\theta, dS)$$

then

$$\theta'_j - \theta_{0j} = e_j^T \Gamma_2 D_{n,2}^{-1}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)) + O_p(\epsilon_n^2).$$

An important consequence of Theorem 2.4.0.1 is that the oracle post-processing, i.e. the post processing associated to B^*, leads to a posterior contraction rate of order $\max(v_{n,k_0+1}, \epsilon_n^2)$. In comparison, Vanilla ABC leads to a posterior contraction rate of order $\max(v_{n,k_0+1}, \epsilon_n)$. The linear post-processing thus corresponds to what would be obtained if ϵ_n was replaced by ϵ_n^2, without increasing the order of the computational cost . This shows the importance of the post-processing as a tool towards dimension reduction, and interestingly the local linear approach already leads to a significant theoretical improvement, even in the general and more realistic framework of summary statistics which have different concentration properties. The proof of Theorem 2.4.0.1 is provided in Section 2.7.4.

2.5 Simulation study

2.5.1 Simulations with reject ABC

We perform accept/reject ABC to obtain Monte Carlo samples from the ABC posterior of Example 1. All of the assumptions of Section 2.2 are satisfied, and so, by Theorem 2.3.0.2, the asymptotic shape of the ABC posterior is available in closed form. The goal of this simulation study is to provide empirical support to this theoretical result.

We recall the data distribution, the prior distribution, and the summary statistics used in Example 1. Data are distributed according to a continuous unit uniform distribution with unknown location parameter $\theta_0 \in \mathbb{R} : y_i \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2}) \;\; \forall i \in \{1, 2, \dots, n\}$. We put a uniform prior on the parameter: $\theta \sim U(0, 1)$. We use the following vector of summary statistics.

$$\eta(y) = \left(y_1, \dots, y_{k_0}, \left(\max_{(k_0+1) \le i \le n} y_i + \min_{(k_0+1) \le i \le n} y_i \right) /2 \right)$$

We have $b(\theta)_i = 0$ and $v_{ni} = 1$ for all $1 \le i \le k_0$ and we have $b(\theta)_{k_0+1} = \theta$ and $v_{n(k_0+1)} = n$. We set the tolerance to be $\epsilon_n = \frac{C}{\sqrt{n}}$ where C is some constant. We are then in the regime where $\lim_{n\to\infty} v_{ni}\epsilon_n = 0 \;\; \forall i \le k_0$, and $\lim_{n\to\infty} v_{n(k_0+1)}\epsilon_n = \infty$. We use the Euclidean norm for distances.

By Lemma 2.7.1.3 we have that this choice for the sequence ϵ_n is equivalent to setting the sequence of ABC acceptance probabilities to be as follows.

$$\alpha_n \propto L_n \epsilon_n^d \propto \epsilon_n^{k_0+d} \propto n^{-\frac{k_0+1}{2}} \tag{2.7}$$

By Theorem 2.3.0.2, we have the following closed-form expression for the ABC posterior for this example, where C is the constant which satisfies $\epsilon_n = \frac{C}{\sqrt{n}}$.

$$\pi_{\epsilon_n}(\theta|\eta(y)) \propto \mathbb{1}_{\{\|\theta-\theta_0\|<\frac{C}{\sqrt{n}}\}} \left(1 - \frac{n\left\|\theta-\theta_0\right\|^2}{C^2}\right)^{\frac{k_0}{2}} \tag{2.8}$$

We perform 3 experiments, one for each of $k_0 = 1, 5, 10$, and let n take values in the range $\{10^5, 2(10^5), 3(10^5), 4(10^5), 5(10^5)\}$. In each case, we ensure that $\alpha_n = 0.3$ for $n = 10^5$. When n increases, α_n decreases accordingly.

Figure 2.1 (left) shows the shapes of the empirical ABC posterior distributions for $k_0 = 1$. Different coloured curves indicate different values of n. Figure 2.1 (right) shows the shapes of the corresponding theoretical ABC posterior distributions of (2.8). Figures 2.2 and 2.3 show the same as the above, with $k_0 = 5$ and $k_0 = 10$ respectively. The similarity between the shapes corresponding to the simulated and to the theoretical ABC posterior curves supports Theorem 2.3.0.2.

The reason why we chose $\alpha_n = 0.3$ for $n = 10^5$ regardless of k_0 was to keep the sizes of reference tables reasonable. If, alternatively, ϵ_n for $n = 10^5$ were kept constant across different choices for k_0, acceptance rates for large k_0 would be much smaller than acceptance rates for large k_0. Low acceptance rates are costly in terms of computational time.

It may seem counter intuitive that the ABC posterior for $k_0 = 1$ is more concentrated than the ABC posterior for $k_0 = 5$ and for $k_0 = 10$ for fixed n. We emphasise that a fixed acceptance probability implies smaller ϵ_n for small k_0 and larger ϵ_n for large k_0. ABC posteriors must therefore not be compared accross different values of k_0 in this study.

2.5.2 Simulations with postprocessing

We return to the example of estimating the location parameter of uniform observations. The data are iid $X_i \sim \mathcal{U}(\theta, \theta+1)$ distribution; we observe $n = 10^4$ realizations. This time, we consider the statistics $S_4 = \frac{1}{\sqrt{n}} \sum_{i=1}^{\sqrt{n}} X_i$ and $S_5 = \min_{1 \le i \le n} X_i$. The

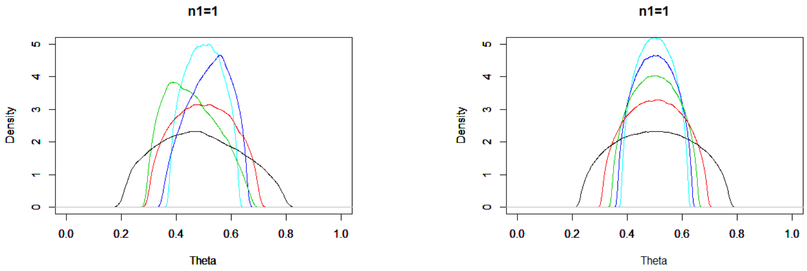


Figure 2.1: Simulated (left) and theoretical (right) ABC posterior for $k_0 = 1$ based on 10000 simulations. Black, red, green, dark blue, and light blue curves correspond to $n = 10^5$, $n = 2(10^5)$, $n = 3(10^5)$, $n = 4(10^5)$, and $n = 5(10^5)$, respectively.

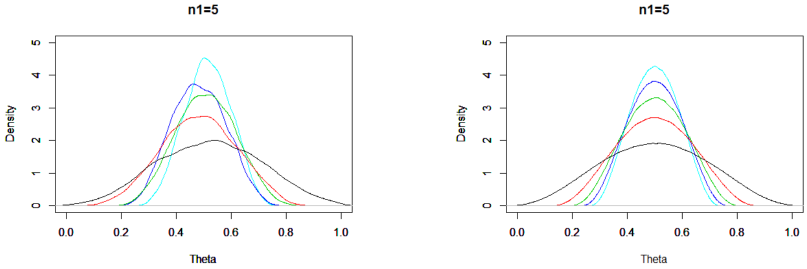


Figure 2.2: Simulated (left) and theoretical (right) ABC posterior for $k_0 = 5$ based on 10000 simulations. Black, red, green, dark blue, and light blue curves correspond to $n = 10^5$, $n = 2(10^5)$, $n = 3(10^5)$, $n = 4(10^5)$, and $n = 5(10^5)$, respectively.

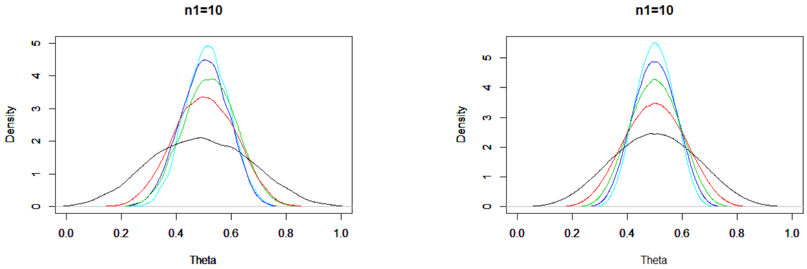


Figure 2.3: Simulated (left) and theoretical (right) ABC posterior for $k_0 = 10$ based on 10000 simulations. Black, red, green, dark blue, and light blue curves correspond to $n = 10^5$, $n = 2(10^5)$, $n = 3(10^5)$, $n = 4(10^5)$, and $n = 5(10^5)$, respectively.

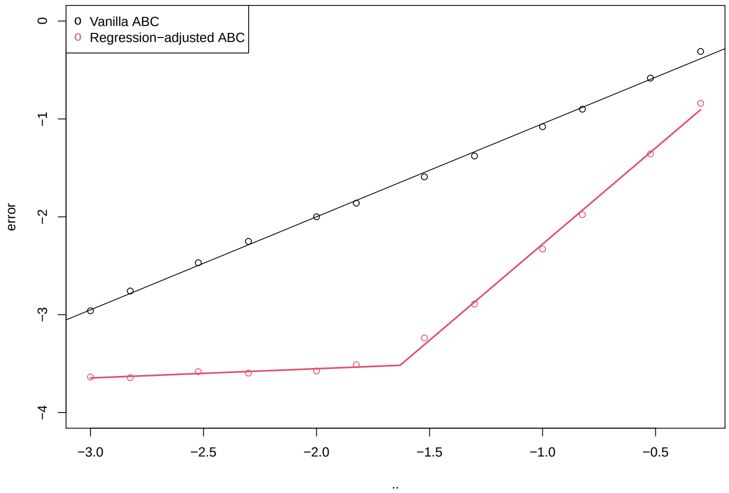


Figure 2.4: Posterior risk for various values of ϵ for the example of Section 2.5.2. Both ϵ and the posterior risk are shown on the log scale.

convergence rates are thus $n^{-1/4}$ for S_4 and n^{-1} for S_5.

We compute the posterior risk $E[(\theta - \theta_0)^2]^{1/2}$, where θ is drawn from the pseudo-posterior for decreasing values of ϵ_n both without and with post-processing. Recall that we expect the risk to decrease at rate ϵ_n in the Vanilla ABC case, but at rate ϵ_n^2 with the post-processing, until the risk reaches a plateau when ϵ_n becomes smaller than $v_{n,k_0+1} = n^{-1}$.

Figure 2.4 shows the posterior risk on the log-log scale (the log is in base 10). Note that as expected, the posterior risk decreases when ϵ decreases. For the Vanilla ABC, the plateau is never reached for computational reasons. A linear regression estimates that in this example, the risk decreases at rate ϵ^γ with $\gamma = 0.95$, very close to the theoretical value of $\gamma = 1$. For ABC with post-processing, segmented regression (as implemented in the R package `segmented` Muggeo (2003)) estimates that the risk decreases at rate ϵ^γ with $\gamma = 1.87$, again close to the theoretical value of $\gamma = 2$. With post-processing, the plateau is reached for $\epsilon \approx 10^{-1.6}$: there is thus no point in decreasing ϵ beyond this value, as we would lose Monte Carlo accuracy but not improve the accuracy of the pseudo-posterior.

2.6 Discussion

We prove posterior consistency, and provide a closed-form expression for the shape of the asymptotic ABC posterior distribution. Unlike in previous work, our results apply to the general case where different components of the summary statistics converge at different rates. In particular, we cover the case where certain components of the summary statistics do not converge at all. This set-up corresponds well to practical situations in applied statistics where large numbers of statistics are used, with potentially varying convergence rates. We also show, under our very general setting that a local linear post-processing approach can lead to significantly faster contraction rates of the pseudo-posterior.

Our theoretical proofs provide, as a byproduct, insight into the effect summary statistic choice and parameter dimension have on the Monte Carlo error. By Lemma 2.7.1.3, acceptance probability is directly proportional to the sequence L_n. As mentioned in Remark 1, L_n will typically take the form $L_n = \prod_{i=1}^{k_0} \epsilon_n v_{ni}$. Thus, we will typically have that the greater the number of slow summary statistics, the faster the acceptance probability will shrink to zero, and so the greater the Monte Carlo error will be. Lemma 2.7.1.3 also illustrates the curse of dimensionality, with acceptance probability decreasing rapidly for large parameter dimension d. This observation is consistent with previous work (Fearnhead and Prangle, 2012)) which suggests making different estimations of subvectors of the vector of parameters separately.

In order for our results to hold true, at least d summary statistics must be used that converge at the fast rate (Assumption 2). Interestingly, adding additional fast converging statistics (i.e. $k - k_0 > d$) will neither change the shape of the asymptotic ABC posterior nor increase the Monte Carlo error.

Throughout our proofs, we make the strong assumption that $\nabla b_{(2)}(\theta_0)$ is of full rank, i.e. sufficiently many summary statistics converge at the fast rate relative to the tolerance (see Assumption 2). Asymptotic results on the ABC posterior in a more general setting where this assumption is lifted will be left to future research.

The local-linear model is one of many post-processing methods for ABC posteriors. An interesting future line of research would be to assess the asymptotic properties of other post-processing methods, for example, the nonlinear conditional heteroscedastic model (Blum and François (2009)).

2.7 Appendix

2.7.1 Statements of lemmas

We consider the following sets: Let $\|\theta - \theta_0\| \leq \lambda_n$ and define $w_n(\theta) = \|\nabla b_2(\theta_0)(\theta - \theta_0)/\epsilon_n\|^2$ together with

$$\begin{aligned}
&A_n(\theta) := \left\{z; \left\|\eta_{(2)}(z) - b_{(2)}(\theta)\right\| \leq \delta_n \epsilon_n (1 \vee \sqrt{w_n(\theta)})/4\right\} \\
&E_n(\theta) := \{z; \|\eta(y) - \eta(z)\| \leq \epsilon_n\}; \quad \tilde{E}_n := \{z; \left\|\eta_{(1)}(y) - \eta_{(1)}(z)\right\| \leq \epsilon_n\} \\
&E'_n(\theta) := \{z; \left\|\eta_{(1)}(y) - \eta_{(1)}(z)\right\| \leq \epsilon_n \left(1 - w_n(\theta) - \delta_n\right)^{\frac{1}{2}}\} \quad \text{if} \quad w_n(\theta) < 1 - \delta_n \\
&E''_n(\theta) := \{z; \left\|\eta_{(1)}(y) - \eta_{(1)}(z)\right\| \leq \epsilon_n \left(1 - w_n(\theta) + \delta_n\right)^{\frac{1}{2}}\}
\end{aligned}$$

Obviously $\tilde{E}_n(\theta) \subset E_n(\theta)$ and $E'_n(\theta) \subset E_n"(\theta) \subset \tilde{E}_n(\theta)$ where the last inequality holds if $\delta_n < w_n(\theta) < 1 - \delta_n$.

Lemma 2.7.1.1. *We can choose $\lambda_n, \delta_n = o(1)$ such that the following inequalities hold: for all $M > 0$, $y \in \Omega_n(M)$,*

1. *if $w_n(\theta) \geq 1 + \delta_n$ and $\|\theta - \theta_0\| \leq \lambda_n$*

$$E_n(\theta) = E_n(\theta) \cap A_n(\theta)^c.$$

2. *if $w_n(\theta) \leq M_1$ for $M_1 > 0$,*

$$P_\theta(E'_n(\theta) \cap A_n(\theta)) \leq P_\theta(E_n(\theta) \cap A_n(\theta)) \leq P_\theta(E''_n(\theta) \cap A_n(\theta))$$

Lemma 2.7.1.2. *Let $M_1, M > 0$, we have for $y \in \Omega_n(M)$,*

$$\sup_{\|\theta - \theta_0\| \leq M_1/v_{n,k_0}} \frac{P_\theta(\tilde{E}_n(\theta))}{L_n} = O(1).$$

Lemma 2.7.1.3. *For a given bandwidth ϵ_n, where $\epsilon_n < \lambda_n$, the average probability of accepting in our accept/reject ABC algorithm, α_n is as follows.*

$$\alpha_n = \int \pi(\theta) P_\theta \left(\|\eta(y) - \eta(z)\| \leq \epsilon_n \right) d\theta \asymp L_n \epsilon_n^d$$

2.7.2 Proof of Theorem 2.3.0.1

Let M_0 be an arbitrarily large constant, $\bar{\lambda}_n \geq M_0 \epsilon_n$ be a sequence going to 0 and such that $(\bar{\lambda}_n v_{n,k_0+1})^{-\kappa} = o(L_n \epsilon_n^d)$, with L_n defined in Assumption 3 and κ in Assumption 4. Note that such a $\bar{\lambda}_n$ exists since $v_{n,k_0+1}^{-\kappa} = o(L_n \epsilon_n^d)$. Consider the event

$$\Omega_n(M) = \{y; \|Z_n(\theta_0)\| \leq M\}, \quad Z_n(\theta_0) = D_n(\eta(y) - b(\theta_0)).$$

For all $\epsilon > 0$, there exists $M_\epsilon > 0$ such that $P_0(\Omega_n(M_\epsilon)^c) \leq \epsilon$. We fix ϵ and consider $M = M_\epsilon$. Hereafter we consider $y \in \Omega_n(M)$.

$$\int_{\left\|b_{(2)}(\theta) - b_{(2)}(\theta_0)\right\| \geq 2\bar{\lambda}_n} \pi_{\epsilon_n}\left(\theta|\eta(y)\right) d\theta = \frac{\int_{\left\|b_{(2)}(\theta) - b_{(2)}(\theta_0)\right\| \geq 2\bar{\lambda}_n} \pi(\theta) P_\theta \left(\|\eta(y) - \eta(z)\| \leq \epsilon_n \right) d\theta}{\int \pi(\theta) P_\theta \left(\|\eta(y) - \eta(z)\| \leq \epsilon_n \right) d\theta}. \tag{2.9}$$

We first consider the numerator of (2.9). Decomposing, we can see that

$$\begin{aligned}\|\eta(y) - \eta(z)\| \geq \left\|\eta_{(2)}(y) - \eta_{(2)}(z)\right\| &\geq \left\|b_{(2)}(\theta) - b_{(2)}(\theta_0)\right\| \\ &\quad - \left(\left\|\eta_{(2)}(y) - b_{(2)}(\theta_0)\right\| + \left\|\eta_{(2)}(z) - b_{(2)}(\theta)\right\| \right).\end{aligned}$$

Recall that $v_{n,j}\epsilon_n \to \infty$ for $j > k_0$, and when θ belongs to the set $\{\left\|b_{(2)}(\theta) - b_{(2)}(\theta_0)\right\| \geq$

$2\bar{\lambda}_n\}$. Putting these together, we have on $\Omega_n(M)$:

$$\|\eta(y)-\eta(z)\| \leq \epsilon_n \Rightarrow \left(\left\|\eta_{(2)}(y)-b_{(2)}(\theta_0)\right\| + \left\|\eta_{(2)}(z)-b_{(2)}(\theta)\right\|\right) \geq 2\bar{\lambda}_n - \epsilon_n$$

which in turns implies that

$$\left\|\eta_{(2)}(y)-b_{(2)}(\theta_0)\right\| \geq 2(\bar{\lambda}_n-\epsilon_n) \geq 2(1-1/M_0)\bar{\lambda}_n.$$

Thus as soon as $M_0 > 2$

$$\begin{aligned}&\int_{\left\|b_{(2)}(\theta)-b_{(2)}(\theta_0)\right\|\geq 2\bar{\lambda}_n} \pi(\theta)P_\theta\left(\|\eta(y)-\eta(z)\|\leq\epsilon_n\right)d\theta \leq \int \pi(\theta)P_\theta\left(\left\|b_{(2)}(\theta)-\eta_{(2)}(z)\right\| > \bar{\lambda}_n\right)d\theta \\ &\leq (k-k_0)\frac{1}{(\bar{\lambda}_n v_{n,k_0+1})^\kappa}\int\pi(\theta)c(\theta)\pi(\theta)d\theta = o(L_n\epsilon_n^d),\end{aligned} \tag{2.10}$$

where the final inequality above comes from Assumption 4.

To lower bound the denominator of (2.9), we simply apply Lemma 2.7.1.3:

$$\int \pi(\theta)P_\theta\left(\|\eta(y)-\eta(z)\|\leq\epsilon_n\right)d\theta \geq CL_n\epsilon_n^d. \tag{2.11}$$

Going back to (2.9), applying (2.10) and (2.11) we find that

$$\int_{\left\|b_{(2)}(\theta)-b_{(2)}(\theta_0)\right\|\geq 2\bar{\lambda}_n} \pi_{\epsilon_n}\left(\theta|\eta(y)\right)d\theta = o(1).$$

By Assumption 2, the transformation $b(\cdot)_{(2)}$ is bijective, which implies that

$$\int_{\|\theta-\theta_0\|\geq 2a^{\frac{-1}{2}}\bar{\lambda}_n} \pi_{\epsilon_n}\left(\theta|\eta(y)\right)d\theta = o(1),$$

where a is defined to be the largest eigenvalue of the matrix $\nabla b_{(2)}(\theta_0)^T\nabla b_{(2)}(\theta_0)$.

Defining λ_n to be $2a^{\frac{-1}{2}}\bar{\lambda}_n$, we have our result.

2.7.3 Proof of Theorem 2.3.0.2

Proof. The ABC posterior, $\pi_{\epsilon_n}(\theta|\eta(y))$, can be expressed as

$$\pi_{\epsilon_n}(\theta|\eta(y)) = \frac{\pi(\theta)P_\theta(E_n)}{\int_{\mathbb{R}^d}\pi(\theta)P_\theta(E_n)d\theta}, \quad E_n = \{\|\eta(z)-\eta(y)\| \leq \epsilon_n\},$$

We define $m(y,\theta)$ to be $m(y,\theta) = Z_{n,(1)}(\theta_0) + D_{n,(1)}(b_{(1)}(\theta) - b_{(1)}(\theta_0))$. We then can define the quantity $h_n(\theta)$ to be

$$h_n(\theta) := L_n\gamma(m(y,\theta))\mathbb{1}_{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\|\leq\epsilon_n}\left(1 - \frac{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\|^2}{\epsilon_n^2}\right)^{\frac{R}{2}},$$

where $\gamma : \mathbb{R}^{k_0} \to \mathbb{R}^+$, L_n and R are defined in Assumption 3. Note that $m(y,\theta) = Z_{n,(1)}(\theta_0)+D_{n,(1)}\nabla b_{(1)}(\theta-\theta_0)$ satisfies $\|m(y,\theta)\| \leq M+Cv_{n,k_0}\epsilon_n \leq C'$ for some $C, C' > 0$ on $\Omega_n(M)\cap\{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\| \leq \epsilon_n\}$. In particular since $v_{n,k_0}\epsilon_n = o(1)$,

$$m(y,\theta) = Z_{n,(1)}(\theta_0) + o(1). \tag{2.12}$$

Moreover, by Theorem 2.3.0.1, $\pi_{\epsilon_n}(\|\theta-\theta_0\| > \lambda_n|y) = o_{P_0}(1)$ and it is enough to control

$$\Delta_n = \int_{\|\theta-\theta_0\|<\lambda_n}\left|\frac{P_\theta(E_n)}{\int_{\|\theta-\theta_0\|<\lambda_n}P_\theta(E_n)d\theta} - \frac{h_n(\theta)}{\int_{\|\theta-\theta_0\|<\lambda_n}h_n(\theta)d\theta}\right|d\theta.$$

To prove our theorem, it will thus be sufficient to prove that $\Delta_n = o(1)$.

In order to facilitate our demonstration, we define quantities $w_n(\theta)$, $V_n(\theta)$, and $V_n'(\theta)$ as $w_n(\theta) = \frac{\|\nabla b(\theta_0)(\theta-\theta_0)\|^2}{\epsilon_n^2}$, $V_n(\theta) = L_n\mathbb{1}_{\{w_n(\theta)\leq 1\}}(1-w_n(\theta))^{\frac{R}{2}}$, $V_n'(\theta) = L_n\mathbb{1}_{\{w_n(\theta)+\delta_n\leq 1\}}(1-w_n(\theta)-\delta_n)^{\frac{R}{2}}$, where $L_n = o(1)$ is defined in Assumption 3, R is the constant defined in Assumption 3, and where δ_n is an $o(1)$ sequence defined in Lemma 2.7.1.1. The quantity $h_n(\theta)$ may then be expressed more simply as $h_n(\theta) = V_n(\theta)\gamma(m_0(y))$.

We have,

$$\begin{aligned}
\Delta_n &= \int_{\|\theta-\theta_0\|<\lambda_n} \left| \frac{P_\theta(E_n)}{\int_{\|\theta-\theta_0\|<\lambda_n} P_\theta(E_n)d\theta} - \frac{h_n(\theta)}{\int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta)d\theta} \right| d\theta \\
&\leq \int_{\|\theta-\theta_0\|<\lambda_n} P_\theta(E_n) \left| \frac{1}{\int_{\|\theta-\theta_0\|<\lambda_n} P_\theta(E_n)d\theta} - \frac{1}{\int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta)d\theta} \right| + \frac{|P_\theta(E_n) - h_n(\theta)|}{\int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta)d\theta} d\theta \\
&\leq 2\frac{\int_{\|\theta-\theta_0\|<\lambda_n} |P_\theta(E_n) - h_n(\theta)|d\theta}{\int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta)d\theta} := \frac{2N_n}{D_n},
\end{aligned}$$

In order to show that $\Delta_n = o(1)$, we show that $N_n = o(\epsilon_n^d)$ and $D_n \geq C\epsilon_n^d$.

Recall that $\delta_n = o(1)$ slowly. We then split the integral over $|\theta - \theta_0| \leq \lambda_n$ into $w_n(\theta) \leq 1-\zeta$, $1-\zeta \leq w_n(\theta) \leq 1+\sqrt{\delta_n}$ and $w_n(\theta) \geq 1+\sqrt{\delta_n}$, where $\zeta > 0$ is a fixed but arbitrarily small constant. This leads to $N_n \leq N_1 + N_2 + N_3$ with

$$\begin{aligned}
N_1 &= \frac{1}{L_n}\int_{w_n(\theta)\leq 1-\zeta} |P_\theta(E_n) - h_n(\theta)|d\theta \\
N_2 &= \frac{1}{L_n}\int_{1-\zeta\leq w_n(\theta)\leq 1+\delta_n} |P_\theta(E_n) - h_n(\theta)|d\theta \\
N_3 &= \frac{1}{L_n}\int_{w_n(\theta)>1+\delta_n} \mathbb{1}_{\|\theta-\theta_0\|\leq\lambda_n} P_\theta(E_n)d\theta, \qquad (2.13)
\end{aligned}$$

where the reduced integrand of N_3 above comes from the fact that $h_n(\theta) = 0$ when $w_n(\theta) > 1$.

From Lemma 2.7.1.1, $P_\theta(E_n) \leq P_\theta(\tilde{E}_n)$ and using Lemma 2.7.1.2, $P_\theta(\tilde{E}_n) \leq CL_n$ uniformly over $w_n(\theta) \leq 2$. By definition of h_n we also have that $h_n(\theta) \leq L_n\gamma(m(y;\theta))$ and when $w_n(\theta) \leq 2$ and on $\Omega_n(M)$, $\gamma(m(y;\theta) \leq \sup_{\|m\|\leq 2M}\gamma(m) < \infty$ so that h_n/L_n is uniformly bounded. Hence on Ω_n, there exists $C > 0$ such that using the change of variable $u = \nabla b_2(\theta_0)(\theta-\theta_0)$ and using the polar coordinates of u,

$$N_2 \leq C\int \mathbb{1}_{1-\zeta\leq w_n(\theta)\leq 1+\delta_n} d\theta \lesssim C\int_{\epsilon_n(1-\zeta)^{1/2}}^{\epsilon_n(1+deltan_n)^{1/2}} r^{d-1}dr \lesssim \epsilon_n^d\zeta. \qquad (2.14)$$

We now study N_1. Firstly, we make use of the inequalities of Lemma 2.7.1.1 to upper and lower bound the quantity $P_\theta(E_n)$. We have

$$\begin{aligned} P_\theta(E_n) &= P_\theta(E_n \cap A_n) + P_\theta(E_n \cap A_n^c) \leq P_\theta(E_n'' \cap A_n) + P_\theta(\tilde{E}_n \cap A_n^c) \\ &\leq P_\theta(E_n'') + P_\theta(\tilde{E}_n \cap A_n^c), \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} P_\theta(E_n) &\geq P_\theta(E_n \cap A_n) \geq P_\theta(E_n' \cap A_n) \\ &= P_\theta(E_n') - P_\theta(E_n' \cap A_n^c) \geq P_\theta(E_n') - P_\theta(\tilde{E}_n \cap A_n^c). \end{aligned} \tag{2.16}$$

Combining (2.15) and (2.16), and using the triangle inequality, we find

$$|P_\theta(E_n) - h_n(\theta)| \leq \max\left\{|P_\theta(E_n') - h_n(\theta)|, |P_\theta(E_n'') - h_n(\theta)|\right\} + P_\theta(\tilde{E}_n \cap A_n^c).$$

Without loss of generality we assume that $\max\left\{|P_\theta(E_n') - h_n(\theta)|, |P_\theta(E_n'') - h_n(\theta)|\right\} = P_\theta(E_n') - h_n(\theta)|$. It then follows that

$$N_1 \leq \frac{1}{L_n}\int_{w_n(\theta)\leq 1-\zeta} |P_\theta(E_n') - h_n(\theta)| d\theta + \frac{1}{L_n}\int_{w_n(\theta)\leq 1-\zeta} P_\theta(\tilde{E}_n \cap A_n^c) d\theta$$

Now using (2.27),

$$P_\theta(E_n') = P_\theta\left(\sum_{j=1}^{k_0} \frac{[Z_{n,j}(\theta) - m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)]^2}{v_{n,j}^2} \leq \epsilon_n^2(1 - w_n(\theta) - \delta_n)\right)$$

To be able to apply Assumption 3, we need to replace $(1 - w_n(\theta) - \delta_n)$ by a constant t. To do so we consider the slices $S_i = \{\theta : t_i < 1 - w_n(\theta) - \delta_n \leq t_{i+1}\}$, where

$$t_1 = \zeta, \quad t_{i+1} = (1+\zeta)t_i \quad i \leq T_\zeta - 1,$$

and T_ζ is the smallest integer satisfying $\sqrt{\zeta}(1+\zeta)^{T_\zeta} \geq 1$. We note that the union of sets $\cup_{i=1}^{(T_\zeta - 1)} S_i$ covers $w_n(\theta) \leq 1-\zeta$ and that the sequence $\{t_1, \ldots, t_{T_\zeta}\}$ has the following properties.

$$\frac{t_{i+1}}{t_i} = 1 + \zeta = 1 \quad \forall i \in \{1, \dots, (T_\zeta - 1)\} \tag{2.17}$$

$$\frac{t_i}{t_{i+1}} = 1 - \frac{\zeta}{\zeta + 1} = 1 \quad \forall i \in \{1, \dots, (T_\zeta - 1)\}. \tag{2.18}$$

Then, using that for $\theta \in S_i$,

$$P_\theta(E_n') \leq P_\theta\left(\sum_{j=1}^{k_0} \frac{[Z_{n,j}(\theta) - m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)]^2}{v_{n,j}^2} \leq \epsilon_n^2 t_{i+1}\right)$$
$$P_\theta(E_n') \geq P_\theta\left(\sum_{j=1}^{k_0} \frac{[Z_{n,j}(\theta) - m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)]^2}{v_{n,j}^2} \leq \epsilon_n^2 t_i\right)$$

and that using assumption 3, we bound on $\Omega_n(M)$, uniformly over $w_n(\theta) \leq 1 - \zeta$,

$$P_\theta(E_n') \leq L_n t_{i+1}^{R/2}(\gamma(m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)) + o(1)$$
$$P_\theta(E_n') \geq L_n t_i^{R/2}(\gamma(m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)) + o(1)$$

Moreover γ is uniformly continuous over any compact and since $\|m_j(y;\theta) + O(v_{n,k_0}\epsilon_n^2)\| \leq 2M$ for M large enough,

$$P_\theta(E_n') \leq L_n t_{i+1}^{R/2}(\gamma(m_j(y;\theta)) + o(1))$$
$$P_\theta(E_n') \geq L_n t_i^{R/2}(\gamma(m_j(y;\theta)) + o(1))$$

which in turns implies that uniformly over $w_n(\theta) \leq 1 - \zeta$,

$$\begin{aligned}\frac{|P_\theta(E_n') - h_n(\theta)|}{L_n} &\leq |t_{i+1}^{R/2} - t_i^{R/2}|\gamma(m_j(y;\theta)) + o(1) \leq R\zeta(\zeta^{R/2-1} + 1) \sup_{\|m\| \leq 2M} \gamma(m) + o(1) \\ &\lesssim \zeta^{R/2 \wedge 1} + o(1).\end{aligned}$$

We thus have

$$N_1 \leq \frac{1}{L_n} \int_{w_n(\theta) \leq 1-\zeta} P_\theta(\tilde{E}_n \cap A_n^c) d\theta + O(\zeta^{R/2 \wedge 1} \epsilon_n^d).$$

We now study $P_\theta(\tilde{E}_n \cap A_n^c)$. By definition and on $\Omega_n(M) \cap A_n^c$,

$$\|D_{n,(2)}^{-1} Z_{n,(2)}(\theta)\| \geq \frac{\delta_n \epsilon_n}{4} - \frac{M}{v_{n,k_0+1}} \geq \frac{\delta_n \epsilon_n}{8}$$

as soon as $v_{n,k_0+1}\delta_n\epsilon_n$ goes to infinity. Hence

$$P_\theta(\tilde{E}_n \cap A_n^c) \leq P_\theta\left(\left\{\|D_{n,(2)}^{-1} Z_{n,(2)}(\theta)\| \geq \frac{\delta_n \epsilon_n}{8}\right\} \cap \left\{\|\eta(z)_{(1)} - \eta(y)_{(1)}\| \leq \epsilon_n\right\}\right)$$

Using the proof of Lemma 2.7.1.2, when $w_n(\theta) \leq 1$,

$$\eta_{(1)}(z) - \eta_{(1)}(y) = D_{n,1}^{-1}[Z_{n,(1)}(\theta) - m(y;\theta) + O(v_{n,1}\epsilon_n^2)] = D_{n,1}^{-1}[Z_{n,(1)}(\theta) - m(y;\theta)] + o(\epsilon_n)$$

Moreover $m(y;\theta) = Z_{n,(1)}(\theta_0) + O(\epsilon_n)$, so that $\|\eta_{(1)}(z) - \eta_{(1)}(y)\| \leq \epsilon_n$ implies that $\|D_{n,1}^{-1}[Z_{n,(1)}(\theta) - Z_{n,(1)}(\theta_0)]\| \leq M_1\epsilon_n$ for some $M_1 > 0$. We then have using $\|Z_{n,(1)}(\theta_0)\| \leq M$

$$\begin{aligned} P_\theta(\tilde{E}_n \cap A_n^c) &\leq E_\theta\left(\mathbb{1}_{\|D_{n,1}^{-1}[Z_{n,(1)}(\theta) - Z_{n,(1)}(\theta_0)]\| \leq M_1\epsilon_n} P_\theta\left(\|D_{n,(2)}^{-1} Z_{n,(2)}(\theta)\| \geq \frac{\delta_n\epsilon_n}{8} \left| Z_{n,(1)}\right.\right)\right) \\ &\leq \sum_{j=k_0+1}^{k} E_\theta\left(\mathbb{1}_{\|D_{n,1}^{-1}[Z_{n,(1)}(\theta) - Z_{n,(1)}(\theta_0)]\| \leq M_1\epsilon_n} P_\theta\left(|Z_{n,j}(\theta)| \geq \frac{v_{n,j}\delta_n\epsilon_n}{8(k-k_0)} \left| Z_{n,(1)}\right.\right)\right) \\ &\leq (k-k_0)\bar{\rho}_n(v_{n,k_0+1}\delta_n\epsilon_n t_0) P_\theta\left(\|D_{n,1}^{-1}[Z_{n,(1)}(\theta) - Z_{n,(1)}(\theta_0)]\| \leq M_1\epsilon_n\right) = o(L_n) \end{aligned}$$

uniformly in θ, where the last two bounds come from Assumption 5 and Lemma 2.7.1.2. Finally this leads to

$$N_1 = o(\epsilon_n^d).$$

We now control N_3. Recall that

$$N_3 = \frac{1}{L_n} \int_{w_n(\theta)>1+\delta_n} \mathbb{1}_{\|\theta-\theta_0\|\leq\lambda_n} P_\theta(E_n) d\theta.$$

We have, from Lemma 2.7.1.1,

$$N_3 = \frac{1}{L_n} \int_{w_n(\theta)>1+\delta_n} \mathbb{1}_{\|\theta-\theta_0\|\leq\lambda_n} P_\theta(E_n \cap A_n^c) d\theta \leq \frac{1}{L_n} \int_{w_n(\theta)>1+\delta_n} \mathbb{1}_{\|\theta-\theta_0\|\leq\lambda_n} P_\theta(\tilde{E}_n \cap A_n^c) d\theta.$$

We then bound, for $\|\theta - \theta_0\| \leq \lambda_n$ and $y \in \Omega_n(M)$,

$$\begin{aligned}
P_\theta(\tilde{E}_n \cap A_n^c) &= E_\theta \left[\mathbb{1}_{\tilde{E}_n}(Z_{n,(1)}) P_\theta \left(\left\| \eta_{(2)}(z) - b_{(2)}(\theta) \right\| > \delta_n \epsilon_n \sqrt{w_n(\theta)}/4 \middle| Z_{n,(1)} \right) \right] \\
&\leq \sum_{j=k_0+1}^{k} E_\theta \left[\mathbb{1}_{\tilde{E}_n}(Z_{n,(1)}) P_\theta \left(|Z_{n,j}\theta)| > v_{n,j} \delta_n \epsilon_n \sqrt{w_n(\theta)}/4 \middle| Z_{n,(1)} \right) \right] \\
&\leq \sum_{j=k_0+1}^{k} \bar{\rho}(v_{n,j} \delta_n \epsilon_n \sqrt{w_n(\theta)}/4) P_\theta \left[\tilde{E}_n \right]
\end{aligned}$$

since $v_{n,j}\delta_n\epsilon_n\sqrt{w_n(\theta)} \gtrsim v_{n,k_0+1}\delta_n\epsilon_n \to \infty$. This implies that there exists $m_1 > 0$ such that

$$\begin{aligned}
N_{3,1} &:= \frac{1}{L_n} \int_{w_n(\theta)>1+\delta_n} \mathbb{1}_{\|\theta-\theta_0\|\leq M_1/v_{n,k_0}} P_\theta(E_n \cap A_n^c) d\theta \\
&\lesssim \frac{1}{(v_{n,k_0+1}\delta_n)^\kappa} \int_{m_1\epsilon_n \leq \|\theta-\theta_0\| \leq M_1/v_{n,k_0}} \|\theta - \theta_0\|^{-\kappa} P_\theta(\tilde{E}_n) d\theta \\
&= O((L_n(v_{n,k_0+1}\delta_n)^{-\kappa}) \int_{m_1\epsilon_n}^{M_1/v_{n,k_0}} r^{d-1-\kappa} dr \\
&= O(L_n(v_{n,k_0+1}\delta_n\epsilon_n)^{-\kappa}\epsilon_n^d) = o(L_n\epsilon_n^d)
\end{aligned}$$

since $\kappa > d$. We also have

$$\begin{aligned}
N_{3,2} &:= \frac{1}{L_n}\int_{\lambda_n \geq \|\theta-\theta_0\| \leq M_1/v_{n,k_0}} P_\theta(E_n \cap A_n^c)d\theta \\
&\lesssim \frac{1}{(v_{n,k_0+1}\delta_n)^\kappa}\int_{M_1/v_{n,k_0} \leq \|\theta-\theta_0\| \leq \lambda_n} \|\theta-\theta_0\|^{-\kappa} P_\theta(\tilde{E}_n)d\theta \\
&= O((v_{n,k_0+1}\delta_n)^{-\kappa})\int_{M_1/v_{n,k_0}}^{\lambda_n} r^{d-1-\kappa}dr \\
&\lesssim \frac{v_{n,k_0}^{-d+\kappa}}{(v_{n,k_0+1}\delta_n)^\kappa} \lesssim \epsilon_n^d \delta_n^{-\kappa}\left(\frac{v_{n,k_0}}{v_{n,k_0+1}}\right)^{\kappa-d}(\epsilon_n v_{n,k_0+1})^{-d} = o(L_n \epsilon_n^d)
\end{aligned}$$

by assumption on $v_{n,k_0}, v_{n,k_0+1}, \epsilon_n$.

We now consider the order of D_n.

$$\begin{aligned}
D_n &:= \frac{1}{L_n}\int_{\|\theta-\theta_0\| < \lambda_n} h_n(\theta)d\theta \\
&= \int_{\|\theta-\theta_0\| < \lambda_n} \mathbb{1}_{\{w_n(\theta)\leq 1\}}(1-w_n(\theta))^{\frac{R}{2}}\gamma\left(\eta_{(1)}(y) - b_{(1)}(\theta_0)\right)d\theta \\
&\geq C\int_{\|\theta-\theta_0\| < \lambda_n} \mathbb{1}_{\{w_n(\theta)\leq 1\}}(1-w_n(\theta))^{\frac{R}{2}}d\theta \\
&\geq C\int_{\frac{1}{2}\leq w_n(\theta)\leq 1} \frac{1}{2}d\theta \\
&= (C+o(1))\epsilon_n^d. \qquad (2.19)
\end{aligned}$$

The third line of the set of equations above comes from the fact that $\gamma(\cdot)$ is lower bounded by a positive constant.

Combining the upper bound on N_n and (2.19), we have

$$\Delta_n = \frac{N_n}{D_n} = \frac{o(\epsilon_n^d)}{C(\epsilon_n^d)} = o(1). \qquad (2.20)$$

We thus have that the ABC posterior, $\pi_{\epsilon_n}(\theta)$, converges in distribution to

$$\frac{h_n(\theta)}{\int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta)d\theta} \propto \mathbf{1}_{\{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\| \leq \epsilon_n\}}\left(1 - \frac{\left\|\nabla b_{(2)}(\theta_0)(\theta-\theta_0)\right\|^2}{\epsilon_n^2}\right)^{\frac{R}{2}}$$

as wanted.

□

2.7.4 Proof of Theorem 2.4.0.1

Proof. Let $m_0 = E_{\theta_0}(Z_n(\theta_0))$. Minimizing in B, β_0 , $L(\beta, \beta_0)$ is equivalent to minimizing in B

$$E_n[\|\theta - E_n(\theta) - B\tilde{S}\|^2] = \sum_{j=1}^{d} E_n[(\tilde{\theta}_j - \beta(j)^T \tilde{S}]^2],$$

which we write $\sum_{j=1}^{d} L_j(\beta(j))$ and where $\beta(j)$ is the j-th row of B, $\tilde{\theta} = \theta - E_n(\theta)$ and $\tilde{S} = S - E_n(S)$. We can thus study the terms L_j separately. Let $j \leq d$, we have

$$L_j(\beta(j)) = V_n(\theta_j) + \beta(j)^T E_n(\tilde{S}\tilde{S}^T)\beta(j) - 2\beta(j)^T E_n(\tilde{S}\tilde{\theta}_j)$$

$$\tilde{S} = \epsilon_n(\nabla b(\theta_0)u + \epsilon_n R(u)) + D_n^{-1}[Z_n(\theta) - E_n(Z_\theta)] \tag{2.21}$$

where $\epsilon_n^2 R(u) = b(\theta(u)) - E_n(b(\theta(u)) - \epsilon_n \nabla b(\theta_0)u = O(\epsilon_n^2 \|u\|^2)$. We first study $E_n(\tilde{S}\tilde{S}^T)$. First note that

$$E_n(Z_\theta) = E_n(E_\theta(Z_n(\theta)) = E_{\theta_0}(Z_n(\theta_0) + O(\|\theta - \theta_0\|) = m_0 + O(\|\theta - \theta_0\|). \tag{2.22}$$

Then writing $\tilde{Z}_n(\theta) = Z_n(\theta) - E_n(Z_n(\theta))$,

$$\begin{aligned} E_n(\tilde{S}\tilde{S}^T) &= \epsilon_n^2 \nabla b(\theta_0) E_n(uu^T) \nabla b(\theta_0)^T + D_n^{-1} \tilde{Z}_n(\theta) \tilde{Z}_n(\theta)^T D_n^{-1} \\ &\quad + 2\epsilon_n D_n^{-1} E_n(\tilde{Z}_n(\theta) u^T) \nabla b(\theta_0)^T + 2\epsilon_n^2 D_n^{-1} E_n(\tilde{Z}_n(\theta) R(u)^T) + O(\epsilon_n^3 E_n(\|u\| \|R(u)\|) \end{aligned}$$

Since for any function $H(u)$, using $\tilde{Z}_n(\theta) = Z_n(\theta) - E_\theta(Z_n(\theta)) + E_\theta(Z_n(\theta)) - E_n(Z_n(\theta))$ together with (2.22)

$$E_n(\tilde{Z}_n(\theta) H(u)) = E_n([E_\theta(Z_n(\theta)) - E_n(Z_n(\theta))] H(u)) = O(\epsilon_n E_n(\|u\| |H(u)|),$$

we obtain that

$$\begin{aligned} E_n(\tilde{S}\tilde{S}^T) &= \epsilon_n^2 \nabla b(\theta_0) E_n(uu^T) \nabla b(\theta_0)^T + D_n^{-1} \tilde{Z}_n(\theta) \tilde{Z}_n(\theta)^T D_n^{-1} \\ &\quad + O\left(\epsilon_n^2 \frac{E_n(\|u\|^3|)}{v_{n,1}}\right) \end{aligned} \tag{2.23}$$

We write $\tilde{S}_{(1)} = (\tilde{S}_1, \ldots, \tilde{S}_{k_0})$ and $\tilde{S}_{(2)} = \tilde{S}_{k_0+1}, \ldots, \tilde{S}_J$.

$$\begin{aligned} E_n(\tilde{S}_{(2)}\tilde{S}_{(2)}^T) &= \epsilon_n^2 \nabla b_2(\theta_0) E_n(uu^T) \nabla b_2(\theta_0)^T + O\left(\frac{\epsilon_n^2}{v_{n,1}} + \frac{1}{v_{n,k_0+1}^2}\right) \\ &= \epsilon_n^2 \nabla b_2(\theta_0) E_n(uu^T) \nabla b_2(\theta_0)^T + o(\epsilon_n^2), \end{aligned}$$

and

$$E_n[\tilde{S}_{(1)}\tilde{S}_{(1)}^T] = D_{n,1}^{-1} C_{1,1} D_{n,1}^{-1} + \epsilon_n^2 \nabla b_1(\theta_0) E_n(uu^T) \nabla b_1(\theta_0)^T + O(\epsilon_n^3) + O(\frac{\epsilon_n^2}{v_{n,1}})$$

$$E_n(\tilde{S}_{(2)}\tilde{S}_{(1)}^T) = \epsilon_n^2 \nabla b_2(\theta_0) E_n(uu^T) \nabla b_1(\theta_0)^T + D_{n,2}^{-1} C_{2,1} D_{n,1}^{-1} + O\left(\frac{\epsilon_n^2}{v_{n,1}}\right)$$

Also

$$\begin{aligned} E_n(\tilde{S}\tilde{\theta}_j) &= \epsilon_n^2 \nabla b(\theta_0) E_n(uu_j) + \epsilon_n D_n^{-1} E_n(\tilde{Z}_n(\theta) u_j) + O\left(\epsilon_n^3\right) \\ &= \epsilon_n^2 \nabla b(\theta_0) E_n(uu_j) + \epsilon_n D_n^{-1} O(\epsilon_n) + O\left(\epsilon_n^3\right) \end{aligned} \tag{2.24}$$

Finally we obtain that

$$\begin{aligned} L_j(\beta(j)) &= V_n(\theta_j) + \beta_{(1)}(j)^T [D_{n,1}^{-1} C_{1,1} D_{n,1}^{-1} + \epsilon_n^2 \nabla b_1(\theta_0) E_n(uu^T) \nabla b_1(\theta_0)^T + O(\epsilon_n^2/v_{n,1})] \beta_{(1)}(j) \\ &\quad + \epsilon_n^2 \beta_{(2)}(j)^T \nabla b_2(\theta_0) E_n(uu^T) \nabla b_2(\theta_0)^T \beta_{(2)}(j) + 2\beta_{(1)}(j)^T D_{n,1}^{-1} C_{1,2} D_{n,2}^{-1} \beta_{(2)}(j) \\ &\quad - 2\epsilon_n^2 \beta(j)^T \nabla b(\theta_0) E_n(uu_j) + O\left(\epsilon_n^3\right) + O(\epsilon_n^2/v_{n,1}) \end{aligned}$$

with $C = E_n(\tilde{Z}_n(\theta)\tilde{Z}_n(\theta)^T)$, $C_{1,1}$ is the top left submatrix of dimension k_0, $C_{2,2}$ the bottom right with dimension $k - k_0$ and $C_{1,2}$ the top right with dimensions $k_0, k - k_0$.

Note that $\nabla b_2(\theta_0)E_n(uu^T)\nabla b_2(\theta_0) = \nabla b_2(\theta_0)[E_h(uu^T) + o(1)]\nabla b_2(\theta_0)$, is positive semi-definite where

$$E_h(uu^T) = \frac{\int_{B_2} uu^T(1 - \|\nabla b_2(\theta_0)u\|^2)^{R/2}du}{\int_{B_2}(1 - \|\nabla b_2(\theta_0)u\|^2)^{R/2}du}, \quad B_2 = \{u \in \mathbb{R}^d;\ \|\nabla b_2(\theta_0)u\| \leq 1\}.$$

Minimizing $L_j(\beta(j))$ boils down to minimizing in $\tilde{\beta}_2 = \beta_{(2)}(j), \tilde{\beta}_1 = D_{n,1}^{-1}\beta_{(1)}(j)/\epsilon_n$

$$\begin{aligned}\tilde{L}(\tilde{\beta}) &= \tilde{\beta}_1^T[C_{1,1} + D_{n,1}^{-1}\nabla b_1(\theta_0)E_n(uu^T)\nabla b_1(\theta_0)^T D_{n,1}^{-1} + O(1/v_{n,1}^3)]\tilde{\beta}_1 \\ &+ \beta_2^T\nabla b_2(\theta_0)E_n(uu^T)\nabla b_2(\theta_0)^T\beta_2 + 2\epsilon_n^{-1}\beta_1 C_{1,2}D_{n,2}^{-1}\beta_2 - 2\beta_2^T\nabla b_2(\theta_0)E_n(uu_j) \\ &- 2\epsilon_n\beta_1^T D_{n,1}\nabla b_2(\theta_0)E_n(uu_j) + O(1/v_{n,1}) \\ &= \tilde{\beta}_1^T C_{1,1}\tilde{\beta}_1^T + \beta_2^T\nabla b(\theta_0)E_n(uu^T)\nabla b_2(\theta_0)^T\beta_2 - 2\beta_2^T\nabla b_2(\theta_0)E_n(uu_j) + O(1/v_{n,1})\end{aligned}$$

Any minimum verifies

$$C_{11}\tilde{\beta}_1 = O(1/v_{n,1}), \quad \nabla b_2(\theta_0)^T\beta_2 = E_n(uu^T)^{-1}E_n(uu_j) + O(1/v_{n,1})$$

In particular the minimum with smaller norm satisfies at the limit

$$\beta_1^* = 0, \quad \nabla b_2(\theta_0)^T\beta_2^* = E_n(uu^T)^{-1}E_n(uu_j) = e_j$$

which is the j-th vector in the canonical bases of $\mathbb{R}^d$. This proves the first part of Theorem 2.4.0.1. We now study $\theta' - \theta_0 = \theta - \theta_0 - B^*(S - S_0)$.

We have for all $j \leq d$

$$\begin{aligned}\theta_j' - \theta_{0j} &= \theta_j - \theta_{0j} - (\beta_2^*(j))^T\nabla b_2(\theta_0)(\theta - \theta_0) + \beta_2^*(j)^T D_{n,2}^{-1}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)) + O(\epsilon_n^2) \\ &= \beta_2^*(j)^T D_{n,2}^{-1}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)) + O(\epsilon_n^2)\end{aligned}$$

□

2.7.5 Proof of Lemma 2.7.1.1

Proof. Throughout this proof C denotes a generic constant whose value is of no importance and can vary from one line to the next.

Let $\delta_n = o(1)$ such that $\delta_n v_{n,k_0+1}\epsilon_n \to \infty$. Let $M > 0$ and consider $y \in \Omega_n(M)$, then for all $\|\theta - \theta_0\| \leq \lambda_n = o(1)$,

$$\begin{aligned}
\left\|\eta_{(2)}(z) - \eta_{(2)}(y)\right\|^2 &= \left\|\nabla_2(\theta_0)(\theta-\theta_0)(1+O(\lambda_n))) + D^{-1}_{n,(2)}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)\right\|^2 \\
&\geq \epsilon_n^2\left(w_n(\theta)(1-C\lambda_n) + \left\|\frac{D^{-1}_{n,(2)}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)}{\epsilon_n}\right\|^2\right. \\
&\quad \left. - 2\sqrt{w_n(\theta)(1-C\lambda_n)}\left\|\frac{D^{-1}_{n,(2)}(Z_{n,(2)}(\theta) - Z_{n,(2)}(\theta_0)}{\epsilon_n}\right\|\right.
\end{aligned}$$

Hence if $w_n(\theta) \geq 1 + \delta_n$, on A_n,

$$\left\|\frac{D^{-1}_{n,(2)}Z_{n,(2)}(\theta)}{\epsilon_n}\right\| \leq \delta_n\sqrt{w_n(\theta)}/4$$

so that

$$\begin{aligned}
\left\|\eta_{(2)}(z) - \eta_{(2)}(y)\right\|^2 &\geq \epsilon_n^2 w_n(\theta)\left(1 - C\lambda_n - \delta_n\sqrt{1+C\lambda_n}/2\right) &(2.25)\\
&\geq \epsilon_n^2(1+\delta_n)(1 - C\lambda_n - \delta_n\sqrt{1+C\lambda_n}/2) > \epsilon_n^2 &(2.26)
\end{aligned}$$

if $w_n(\theta) > 1 + \delta_n$ and as soon as $C\lambda_n < \delta_n/3$ and δ_n is small enough. Hence part 1 of Lemma 2.7.1.1 is proved.

We now prove part 2. Let θ be such that $w_n(\theta) \leq M_1$. We omit θ in the notations $E_n, E'_n, A_n, \tilde{E}_n$. Using the same computations as above , on A_n, if $w_n(\theta) > 1$,

$$\left\|\eta_{(2)}(z) - \eta_{(2)}(y)\right\|^2 \geq \epsilon_n^2 w_n(\theta)\left(1 - C\lambda_n - \delta_n\sqrt{1+C\lambda_n}/2\right) \geq \epsilon_n^2 w_n(\theta)(1-\delta_n)$$

and similarly

$$\left\|\eta_{(2)}(z)-\eta_{(2)}(y)\right\|^2 \leq \epsilon_n^2 w_n(\theta)(1+C\lambda_n)(1+\delta_n/4)^2 \leq \epsilon_n^2 w_n(\theta)(1+\delta_n).$$

Also if $\delta_n \leq w_n(\theta) \leq 1$,

$$\begin{aligned}\left\|\eta_{(2)}(z)-\eta_{(2)}(y)\right\|^2 &\geq \epsilon_n^2\left(\sqrt{w_n(\theta)(1-C\lambda_n)}-\delta_n/4\right)^2\\ \left\|\eta_{(2)}(z)-\eta_{(2)}(y)\right\|^2 &\leq \epsilon_n^2\left(\sqrt{w_n(\theta)(1+C\lambda_n)}+\delta_n/4\right)^2\end{aligned}$$

so that if, $\left\|\eta_{(1)}(z)-\eta_{(1)}(y)\right\|^2 \leq \epsilon_n^2(1-w_n(\theta)-\delta_n)$, $w_n(\theta) \leq 1$ and

$$\|\eta(z)-\eta(y)\|^2 \leq \epsilon_n^2[C\lambda_n+\sqrt{1+C\lambda_n}\delta_n/2-\delta_n \leq \epsilon_n^2$$

by choosing $\lambda_n \leq c\delta_n$ with c small enough. Hence $E_n' \cap A_n \subset E_n \cap A_n$. Similar arguments imply that $E_n \cap A_n \subset E_n" \cap A_n$.

□

2.7.6 Proof of Lemma 2.7.1.2

Proof. We have

$$\eta_{(1)}(z)-\eta_{(1)}(y) = D_{n,1}^{-1}[Z_{n,(1)}(\theta)-Z_{n,(1)}(\theta_0)]+b_{(1)}(\theta)-b_{(1)}(\theta_0)$$

so that if $\|\theta-\theta_0\| \leq v_{n,k_0}^{-1}M_1 \wedge \delta_n$,

$$\eta_{(1)}(z)-\eta_{(1)}(y) = D_{n,1}^{-1}[Z_{n,(1)}(\theta)-m(y;\theta)+O(v_{n,k_0}^{-1}\wedge v_{n,k_0}\delta_n^2)] = D_{n,1}^{-1}[Z_{n,(1)}(\theta)-m(y;\theta)+o(1)]$$

where $m(y;\theta) = Z_{n,(1)}(\theta_0) + D_{n,1}\nabla b_1(\theta_0)(\theta-\theta_0)$. and on $\Omega_n(M)$ $\|m(y;\theta)+o(1)\| \leq 2M$ by choosing M large enough.

$$P_\theta(\tilde{E}_n) = P_\theta\left(\sum_{j=1}^{k_0} \frac{[Z_{n,j}(\theta) - m_j(y;\theta) + o(1)]^2}{v_{n,j}^2} \leq \epsilon_n^2\right)$$

It implies in particular that with K the ball in $\mathbb{R}^{k_0}$ centered at 0 and with radius $2M$, uniformly over $\|\theta-\theta_0\| \leq \lambda_n$,

$$\left|\frac{P_\theta(\tilde{E}_n)}{L_n} - \gamma(m_j(y;\theta)+o(1))\right| \leq \sup_{m\in K}\left|\frac{P_\theta\left(\sum_{j=1}^{k_0}\frac{[Z_{n,j}(\theta)-m_j]^2}{v_{n,j}^2} \leq \epsilon_n^2\right)}{L_n} - \gamma(m)\right| \tag{2.27}$$

$$= o(1), \tag{2.28}$$

where the last equality comes from Assumption 3.

It follows in particular that

$$\sup_{\|\theta-\theta_0\|\leq\lambda_n}\left\|\frac{P_\theta(\tilde{E}_n)}{L_n}\right\| \leq \sup_{m\in K}\gamma(m) + o(1) = O(1).$$

□

2.7.7 Proof of Lemma 2.7.1.3

Proof. We have

$$\begin{aligned}
\alpha_n &= \int \pi(\theta) P_\theta \left(\|\eta(z) - \eta(y)\| < \epsilon_n \right) d\theta \\
&= \int_{\|\theta-\theta_0\|<\lambda_n} h_n(\theta) d\theta + o\left(L_n \epsilon_n^d\right) \\
&= L_n \gamma \left(\eta_{(1)}(y) - b_{(1)}(\theta_0)\right) \int_{\|\theta-\theta_0\|<\lambda_n; w_n(\theta)\leq 1} (1 - w_n(\theta))^{\frac{R}{2}} d\theta + o\left(L_n \epsilon_n^d\right) \\
&= L_n \epsilon_n^d \gamma \left(\eta_{(1)}(y) - b_{(1)}(\theta_0)\right) \det\left(\nabla b_{(2)}(\theta_0)^T \nabla b_{(2)}(\theta_0)\right) \int_{u\leq 1} (1 - u^2)^{\frac{R}{2}} du + o\left(L_n \epsilon_n^d\right) \\
&= L_n \epsilon_n^d \gamma \left(\eta_{(1)}(y) - b_{(1)}(\theta_0)\right) \frac{1}{2} \det\left(\left(\nabla b_{(2)}(\theta_0) \nabla b_{(2)}(\theta_0)\right)^{\frac{1}{2}}\right) \mathrm{Beta}\left(\frac{1}{2}, \frac{R}{2} + 1\right) + o\left(L_n \epsilon_n^d\right) \\
&= (C + o(1)) L_n \epsilon_n^d .
\end{aligned}$$

The third line of the set of equations above comes from (2.20). The fourth line comes from the definition of $h_n(\theta)$. The fifth line comes from a change of variables. □

Chapter 3

A Bayesian non-parametric model for 2-multiple context free grammars

Abstract

The class of context-free grammars is believed to be too restrictive to fully describe all features of natural language. The class of context-sensitive grammars, on the other hand, is too complex to be practical: modelling with them would require an unrealistic amount of computational time. Various mildly context-free grammar formalisms, which may be placed between context-free grammars and context-sensitive grammars in terms of complexity, have thus been proposed in the last few decades. We consider the class of 2-multiple context-free grammars (2-MCFGs) (Seki et al. (1991)), which properly include the class of context-free grammars.

We propose a Bayesian non-parametric model for 2-MCFGs within which a model for context-free grammars is naturally embedded. Our model is inspired by that of Ryder et al. (2023) for context-free grammars, and is based on hierarchical Dirichlet processes. We develop a sequential Monte Carlo algorithm to make inference under this

model. We carry out simulation studies to assess our method.

3.1 Introduction

Informally, a grammar is what defines the structure of a language. It tells us which way words from a language may be joined together in order to form sentences from the language that are valid according to the language's syntax. Grammars were first formalised by Chomsky (1956), who defined them as consisting of the following four components.

- A set $\mathcal{B}$ of nonterminal symbols
- A set $\mathcal{A}$ of terminal symbols
- A set $\mathcal{R}$ of rules, each of the form

$$R : (\mathcal{A} \cup \mathcal{B})^{\star}\mathcal{B}(\mathcal{A} \cup \mathcal{B})^{\star} \rightarrow (\mathcal{A} \cup \mathcal{B})^{\star} \tag{3.1}$$

  where * represents the Kleene star operator, which is defined as the set of strings formed by concatenating one or more of the elements of the set it is applied to.

- A distinguished start nonterminal symbol S where $S \in \mathcal{B}$

The terminal symbols $\mathcal{A}$ are the words of the grammar's language, while the nonterminal symbols and the rules describe in a generative way (as we will explain in later sections) how valid sentences may be formed according to the grammar. The symbol S is used at the beginning of the sentence-generating process. In Section 3.2.1 and Section 3.2.2 we will describe in detail how sentences (strings of terminal symbols) may be formed from two special cases of grammars.

Definition 4. *Given a Chomsky grammar $\mathcal{G} = (\mathcal{B}, \mathcal{A}, \mathcal{R}, S)$, the language $L(\mathcal{G})$ is the set of strings of terminal symbols that can be generated from the grammar $\mathcal{G}$.*

In general, for any grammar, only a subset of the set of possible rules in the above definition will be allowed. The complexity of a grammar is measured by the complexity of the set of rules the grammar allows. Chomsky (1956) classified grammars in terms of the complexity of the rules that they allow. In order of increasing complexity he defined the following four classes of grammars.

$$\text{Regular} \subset \text{Context-free} \subset \text{Context-sensitive} \subset \text{Recursively enumerable}$$

The more complex a grammars model is, the more features of a language it may be able to capture. However, the more complex the grammar model, the more expensive the inference may be in terms of computational time. It is standard for human natural language to be modeled using context-free grammars, and there is a vast literature on efficient algorithms for context-free grammar models (Lari and Young (1990), Johnson et al. (2007)). It is well known, however, that context-free grammars do not capture all features of human natural language. Shieber (1985) demonstrate this for the particular case of the Swiss German language. Furthermore, recent research suggest that the vocalisations of Muriqui monkeys is more complex than context-free (Chatain et al. (2021)).

Since the class of context-sensitive grammars, which comes above the class of context-free grammars in terms of complexity, is considered too complex in practice for simple inference purposes, researchers have proposed intermediate classes of grammars which lie in between context-free and context-sensitive in terms of complexity. Examples of these include head grammars (Pollard (1984)), tree-adjoining grammars (Joshi et al. (1969)), and multiple context-free grammars (Seki et al. (1991)). In this work, we will consider the class of 2-multiple context-free grammars (2-MCFGs), which is properly included in the class of context-sensitive grammars, and which properly includes the class of context-free grammars.

The standard method for making inference on parameters in computational linguistics is based on expectation-maximization techniques, such as the inside-outside algorithm for context-free grammars and its variants (Lari and Young (1990)). In the case of context-free grammars, some researchers have turned to Bayesian methods of inference (Johnson et al. (2007), Goldwater and Griffiths (2007)). These have the advantage of taking into account the uncertainty of all parameters in the framework of a single probabilistic model. When modelling grammars, an important question to consider is the number of latent parameters that describe the structure of the grammar. While including too few parameters corresponds to a lack of expressiveness of the model, including too many parameters can be computationally expensive, unrealistic, and can lead to over-fitting. This choice may be bypassed by putting a nonparametric prior on the space of parameters, which allows the number of latent parameters to be learned adaptively with the data. Nonparametric priors also have the advantage of naturally penalizing grammars with too many parameters.

As we will describe in detail in later sections, each rule in a 2-MCFG can be associated with one nonterminal symbol from the set $\mathcal{B}$. The rules of a 2-MCFG can then be modeled as grouped data, where each group corresponds to the set of rules associated with each of the nonterminal symbols. Dirichlet process mixture models (Antoniak (1974)) are a popular nonparametric model for grouped data. More sophisticated variants of the DP mixture model such as the Dirichlet process hidden Markov model (Beal et al. (2002)) and the hierarchical Dirichlet process (Teh et al. (2006)) (HDP) relax the independence assumption across groups. Liang et al. (2007) and Finkel et al. (2007) have used HDPs to model context-free grammars. While Liang et al. (2007) perform inference by variational Bayes, Finkel et al. (2007) use MCMC techniques to model context-free grammars. In this work, we propose a hierarchical Dirichlet process model for 2-MCFGs and perform inference on parameters by sequential Monte Carlo (SMC). To the best of our knowledge, no previous work has attempted a Bayesian approach to model 2-MCFGs.

The remainder of this chapter is organized as follows. In Section 3.2 we provide some background information, with formal definitions of context-free grammars, 2-multiple context-free grammars, and probabilistic 2-multiple context-free grammars.

In Section 3.3 we describe our Bayesian model in detail and in Section 3.4 we describe our sequential Monte Carlo inference scheme. Results of our simulation studies are presented in Section 3.5, and a discussion is provided in Section 3.6. Details of our SMC scheme are left to appendix Section 3.7.

3.2 Background

As previously mentioned, the main focus of this work is the modelling of the class of 2-multiple context-free grammars, which is an extension of the class of context-free grammars. In this section, we describe in detail the type of rules that are allowed in each of these classes of grammars and show how parse trees and their corresponding sentences can be formed from them. We then introduce the concept of probabilistic grammars, which allows us to model grammars using statistical techniques.

3.2.1 Context-free grammars

There are a number of different ways of formalising the rules of context-free grammar that are weakly equivalent (i.e. that result in exactly the same set of possible sentences). In this work, we express all context-free grammars in Chomsky normal form. All of the rules of a context-free grammar in Chomsky normal form must be written in one of the following two forms.

$$B_j \to B_{k_1} B_{k_2} \tag{3.2}$$

$$B_j \to a_k \tag{3.3}$$

where $B_j, B_{k_1}, B_{k_2} \in \mathcal{B}$ and where $a_k \in \mathcal{A}$.

Rules of the form of Equation 3.2 are called *production rules* and rules of the form of Equation 3.3 are called *emission rules.* For context-free grammars in Chomsky normal form, production rules replace one nonterminal symbol with a pair of nonterminal symbols. Emission rules replace one nonterminal symbol with one terminal symbol.

The structure of a sentence generated from a context-free grammar in Chomsky normal form may be represented by a tree, where the root of the tree is associated with the nonterminal symbol S, where each internal node of the tree is associated with a nonterminal symbol from $\mathcal{B}$ and each leaf of the tree is associated with a terminal symbol from $\mathcal{A}$. Given a context-free grammar, trees are generated as follows. Associate the root node of the tree with the nonterminal symbol S. Pick a rule of the form (3.2) or (3.3) with S on the left-hand-side. If the rule is of the form (3.3), i.e. $S \to a_k$, add a branch stemming to the terminal symbol emitted by the rule. Otherwise, if the rule is of the form (3.2), i.e. $S \to B_{k_1} B_{k_2}$, add two branches stemming to child nodes, and associate each of the child nodes with the nonterminal symbols B_{k_1} and B_{k_2}. Next, select any remaining leaf node of the tree which is associated with a nonterminal symbol, say B_{k_1}. Repeat the above, picking a rule this time with B_{k_1} on the left-hand-side and extending the tree as before. Continue this process until all of the leaf nodes of the tree are associated with terminal symbols.

Given a standard tree as described above, we define its corresponding evaluated tree to be a tree of the same size shape, and structure, but where each node is associated with a string of terminal symbols. An evaluated tree is formed from a standard tree as follows. First, starting from the bottom of the tree, associate all nodes that emit leaves with the single terminal symbol that they emit. Next, associate any internal node that produces a pair of internal nodes with the concatenation of the strings of terminal symbols associated with its two child nodes. For example, suppose that we had a subtree consisting of the following three rules: $B_j \to B_{k_1} B_{k_2}$, $B_{k_1} \to a_{j_1}$ and $B_{k_2} \to a_{j_2}$. The string associated with B_{k_1} in the tree would be a_{j_1}, the string associated with B_{k_2} in the tree would be a_{j_2} and the string associated with B_j in the tree would be $a_{j_1} a_{j_2}$. Carry out this process from the bottom to the top of the tree. The string of terminal symbols associated with the root node of the tree will be the complete sentence that the tree represents.

A sentence is grammatical if there exists a sequence of rules which produce it or

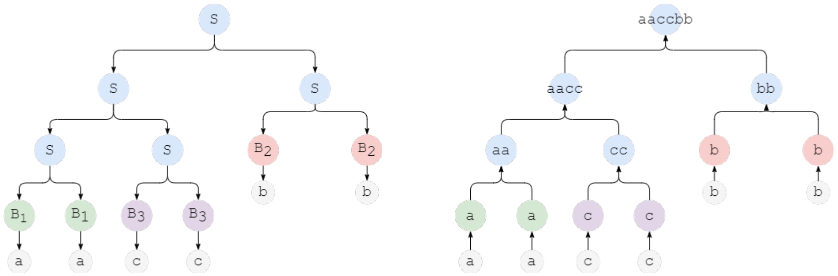


Figure 3.1: Left: A tree generated from the context-free grammar $\mathcal{G}_{D,1}$ (see Example 2). The rules used to generate the tree are (in order): $S \to SS$, $S \to SS$, $, S \to B_1B_1$, $, S \to B_3B_3$, $, S \to B_2B_2$, $B_1 \to a$, $, B_1, \to a$, $B_3 \to c$, $B_3, \to c$, $B_2 \to b$, $, B_2, \to b$. Right: The evaluated version of the context-free tree. At the root of the tree, one reads the sentence *aaccbb*, which is the sentence associated with the tree.

equivalently, a tree which evaluates to it. There may be multiple trees that evaluate to the same grammatical sentence.

Example 2 (Doubles language)**.** *Consider the following context-free grammar.*

$$
\begin{aligned}
\mathcal{G}_{D,1} &= (\mathcal{A}, \mathcal{B}, \mathcal{R}, S) \\
\mathcal{A} &= \{a, b, c\} \\
\mathcal{B} &= \{S, B_1, B_2, B_3\} \\
\mathcal{R} &= \{S \to SS, \quad S \to B_1B_1 \quad S \to B_2B_2, \quad S \to B_3B_3 \\
&\qquad B_1 \to a, \quad B_2 \to b, \quad B_3 \to c\}
\end{aligned}
$$

The grammar $\mathcal{G}_{D,1}$ describes is the language consisting of strings of symbols from $\mathcal{A}$ where each element is repeated twice (for example aabbccaa, bbaabbbbcc or ccaa, etc). We refer to the language produced by $\mathcal{G}_{D,1}$ as the "doubles" language.

$$L(\mathcal{G}_{D,1}) = \{\zeta_1\zeta_1 \cdots \zeta_n\zeta_n | n \in \mathbb{N}, \zeta_i \in \mathcal{A} \ \forall i \in \{1, \ldots, n\}\}$$

3.2.2 2-Multiple context-free grammars (2-MCFGs)

Formally, a 2-MCFG may be represented by a five-tuple $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}, S)$ where $\mathcal{A}, \mathcal{B}$ and S are as in the case of context-free grammars, where $\mathcal{F}$ is some multiset of permutations over 5 elements which we will describe, and where all of the rules in $\mathcal{R}$ must take one of the following two forms.

$$B_j \to f_j[B_{k_1}, B_{k_2}], \qquad f_j \in \mathcal{F} \tag{3.4}$$

$$B_j \to (a_{k_1}, a_{k_2}) \tag{3.5}$$

where $B_j, B_{k_1}, B_{k_2} \in \mathcal{B}$, and where $(a_{k_1}, a_{k_2}) \in \{\mathcal{A} \times \epsilon\} \cup \{\epsilon \times \mathcal{A}\} \cup \mathcal{A}^{\bigotimes 2}$ (where ϵ denotes the empty symbol). Rules of the form of Equation 3.4 are called *production rules* and rules of the form of Equation 3.5 are called *emission rules.* Production rules replace one nonterminal symbol with two nonterminal symbols from $\mathcal{B}$ and one permutation f from $\mathcal{F}$. Emission rules replace one nonterminal symbol either with a pair of terminal symbols from $\mathcal{A}$, or with one terminal symbol from $\mathcal{A}$ and one empty symbol ϵ.

The structure of a sentence generated from a 2-MCFG can be represented by a tree, where the root node is associated with the symbol S, where each internal node is associated with a nonterminal symbol from $\mathcal{B}$ and with a permutation from $\mathcal{F}$, and where leaves of the tree are associated either with elements from $\mathcal{A}$ or the empty symbol ϵ. Given a 2-MCFG, trees are generated as follows. Associate the root node of the tree with the nonterminal symbol S. Pick a rule of the form (3.4) or (3.5) with S on the left-hand-side. If the rule is of the form (3.5), add two branches to the node, one stemming to the symbol a_{k_1} and one stemming to the symbol a_{k_2}. Otherwise, if the rule is of the form (3.4), associate the node with the permutation involved in that rule (f_j), and add two branches to the node, one stemming to the nonterminal symbol B_{k_1} and one stemming to the symbol B_{k_2}. Next, select any remaining leaf node of the tree which is associated with a nonterminal symbol, say B_{k_1}. Repeat the above, picking a rule this time with B_{k_1} on the left-hand side and extending the tree as before. Continue this process until all of the leaf nodes of the tree are associated with terminal symbols.

Recall that in the case of context-free grammars, each node of the evaluated version of a tree is associated with a string of terminal symbols, in such a way that the root node of the evaluated tree is associated with the sentence that the tree represents. In the case of evaluated 2-MCFG trees, each node is associated with a pair of strings of terminal symbols. The sentence that a 2-MCFG tree represents will be the sentence formed by concatenating the two strings associated with the root node of the evaluated version of that tree.

We represent each node of an evaluated 2-MCFG tree by $\langle \zeta_1 | \zeta_2 \rangle$ where ζ_1 and ζ_2 are strings of terminal symbols and where "$|$" is a special symbol that indicates where the string of terminal symbols $\zeta_1\zeta_2$ is split. Whenever an emission rule $B_j \to (a_{k_1}, a_{k_2})$ takes place at some node of the standard tree, we associate that node in the evaluated tree with the pair $\langle a_{k_1} | a_{k_2} \rangle$.

For a production rule $B_j \to [B_{k_1}, B_{k_2}]$, let $\langle \zeta_{1,1} | \zeta_{1,2} \rangle$ and $\langle \zeta_{2,1} | \zeta_{2,2} \rangle$ be the pairs of strings associated with the child nodes corresponding to B_{k_1} and B_{k_2} respectively in the evaluated tree. Recall that $f_j \in \mathcal{F}$ is a permutation over 5 elements. Associate to B_j in the evaluated tree the pair of strings $\langle f_j(\zeta_{1,1}, \zeta_{1,2}, \zeta_{2,1}, \zeta_{2,2}, |) \rangle$. For example, suppose that B_{k_1} were associated with the pair $\langle ab | d \rangle$, that B_{k_2} were associated with the pair $\langle c | \epsilon \rangle$, and that $f_j = (15342)$. Then we would associate B_j with the output of the permutation f_j when applied to the vector $v := (ab, d, c, \epsilon, \text{“}|\text{”}))$. In this example, the output of f_j when applied to v gives $\langle ab | c\epsilon d \rangle$ (and since ϵ refers to the empty symbol, this is equivalent to $\langle ab | cd \rangle$). We refer to this process of associating strings of terminals to the nodes of the tree as evaluating the tree.

Example 3 (Doubles language)**.** *Consider the following 2-MCFG.*

$$\begin{aligned}
\mathcal{G}_{D,2} &= (\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}, S)\\
\mathcal{A} &= \{a, b, c\}\\
\mathcal{B} &= \{S\}\\
\mathcal{F} &= \{f_D = (12534)\}\\
\mathcal{R} &= \{S \to f_D[S, S], \quad S \to (a, a), \quad S \to (b, b), \quad S \to (c, c)\}
\end{aligned}$$

Observe that the grammar only allows one permutation, f_D, which concatenates both strings associated with the first child node and puts them on the left-hand side, and concatenates both strings associated with the second child node and puts them on the right-hand side. As in the grammar $\mathcal{G}_{D,1}$ of Example 2, the grammar $\mathcal{G}_{D,2}$ describes the "doubles" language.

$$L(\mathcal{G}_{D,2}) = \{\zeta_1\zeta_1 \cdots \zeta_n\zeta_n | n \in \mathbb{N}, \zeta_i \in \mathcal{A} \ \forall i \in \{1, \ldots, n\}\}$$

Example 4 (Copy language)**.** *Consider the following 2-MCFG.*

$$\begin{aligned}
\mathcal{G}_{C} &= (\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}, S)\\
\mathcal{A} &= \{a, b, c\}\\
\mathcal{B} &= \{S\}\\
\mathcal{F} &= \{f_C = (13524)\}\\
\mathcal{R} &= \{S \to f_C[S, S], \quad S \to (a, a), \quad S \to (b, b), \quad S \to (c, c)\}
\end{aligned}$$

Observe that the grammar only allows one permutation, f_C, which concatenates the left-hand side string associated with the first child node with the left-hand side string associated with the second child node and puts this on the left-hand side, and concatenates the right-hand side string associated with the first child node with the right-hand side string associated with the second child node, and puts this on the right-hand side.

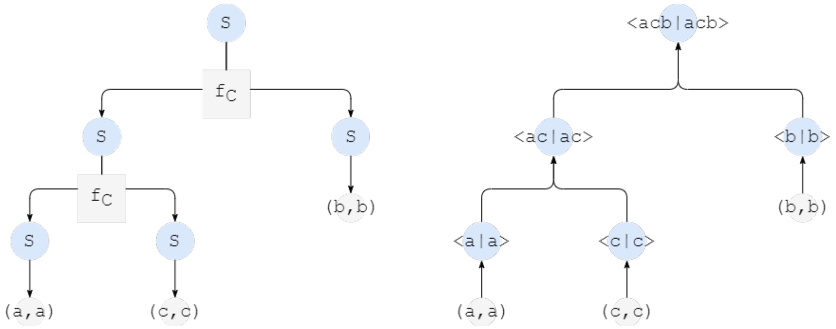


Figure 3.2: Left: A tree generated from the 2-MCFG grammar $\mathcal{G}_C$ (see Example 4). The rules used to generate the tree are (in order): $S \to f_C[S, S]$, $S \to f_C[S, S]$, , $S \to (a, a)$, , $S \to (c, c)$, , $S \to (b, b)$. Right: The evaluated version of the 2-MCFG tree. At the root of the tree, one reads the sentence *acbacb*, which is the sentence associated with the tree.

The grammar $\mathcal{G}_C$ describes the language which consists of two identical strings of terminal symbols concatenated (for example bcbc, abcabc or bccbcc, etc). Indeed, it is straightforward to show recursively that in the evaluated tree, each internal node is associated with a pair $\langle \zeta | \zeta \rangle$ for some $\zeta \in \mathcal{A}^\star$. We refer to the language produced by $\mathcal{G}_C$ as the "copy" language.

$$L(\mathcal{G}_C) = \{\zeta^2 | \zeta \in \mathcal{A}^\star\}$$

Remark 3. *2-multiple context-free grammars are a special case of the more general class of $m-$multiple context-free grammars, where m may be any posistive integer (Seki et al. (1991)). The class of $m-$multiple context-free grammars, for some arbitrary $m \in \mathbb{N}$ may be defined just as the definition for 2-MCFGs, but with each nonterminal symbol being associated with an $m-$tuple of strings of terminal symbols (rather than a pair of strings of nonterminal symbols). Emission rules in $m-$MCFGs emit m terminal symbols (rather than 2 nonterminal symbols). It can be proved that m_1-MCFGs are properly included in m_2-MCFGs for any $m_1 < m_2$ (Seki et al. (1991)).*

Remark 4. *Context-free grammars are equivalent to 1-MCFGs.*

3.2.3 Probabilistic 2-MCFGs

A probabilistic grammar is a grammar that additionally assigns some probability to each of its rules. A probabilistic 2-MCFG may be written as a 6-tuple $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}, S, \mathcal{J})$, where $\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}$ and S are as in the definition for 2-MCFGs, and where $\mathcal{J}$ denotes a collection of probability distributions over the elements of the set $\mathcal{R}$. In 2-MCFGs, each rule can be associated with the nonterminal symbol B_j on the left-hand side of the rule of the form of Equation 3.5 and Equation 3.4. The elements of $\mathcal{J}$ are the distributions over rules associated with each of the nonterminal symbols in $\mathcal{B}$, in such a way that for each element B_j of $\mathcal{B}$, there exists an element $\mathcal{P}_{R_j}$ in $\mathcal{J}$ that is a probability distribution over rules associated with B_j.

Given a probabilistic grammar, we define the probability of any tree, τ generated from that grammar to be the product of the rules associated with the nodes of that tree. We have

$$\mathcal{P}_\tau(\tau) := \prod_{x \in \tau'} \mathcal{P}_{R_{j(x)}} \left(R_{j(x)}^{q(x)} \right)$$

where τ' denotes the set of internal nodes of the tree τ, where $j(x)$ denotes the nonterminal symbol associated with the node x and where $q(x)$ denotes the index of the rule associated with the node x.

For any sentence in a grammar's language there may be one or more distinct trees that can be generated from the grammar that, when evaluated, form it. For probabilistic grammars, the probability of any sentence in the grammar's language is defined to be the sum over the probabilities of with each of the trees that form that sentence.

3.3 Model

As stated in Section 3.2.2, a 2-MCFG consists of the five-tuple $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{R}, S)$. In our setting, the set of terminal symbols, $\mathcal{A}$, and the set of permutations $\mathcal{F}$ are given and finite. The set of nonterminal symbols $\mathcal{B}$ and the set of rules $\mathcal{R}$ on the other hand, are random and countably infinite.

We model $\mathcal{A}$ with the categorical distribution $\mathcal{P}_{\mathcal{A}}$ with parameters $\bar{\mu}_a = (\mu_a^1, \dots \mu_a^{|\mathcal{A}|})$, with $\mu_a^j \geq 0 \ \forall j$ and $\sum_{j=1}^{|\mathcal{A}|} \mu_a^j = 1$ in such a way that $\mathcal{P}_{\mathcal{A}}(a_j) = \mu_a^j$, where a_j represents the j^{th} nonterminal symbol. In a similar way, we model the set of 120 possible permutations over 5 elements with the categorical distribution $\mathcal{P}_{\mathcal{F}}$, with parameters $\bar{\mu}_f = (\mu_f^1, \dots, \mu_f^{120})$.

Let $\mathcal{P}_{\mathcal{B}}$ represent the distribution over the countably infinite set of nonterminal symbols $\mathcal{B}$. We model this distribution with a Dirichlet process, with base measure H_1 and with scaling parameter α_1. We model H_1 with a standard normal distribution (since we will only be interested in its partitions, any continuous distribution will work). We then have, using standard notation

$$\mathcal{P}_{\mathcal{B}} \sim DP\left(\alpha_1, H_1\right), \qquad H_1 = N(0, 1).$$

As described in Section 3.2.2, each rule will be associated with one of the nonterminal symbols (i.e. the nonterminal symbol on the left-hand-side of the expressions of Equation 3.4 and Equation 3.5). To each of the infinitely many nonterminal symbols in the grammar, we assign a distribution over its associated rules. We thus obtain infinitely many rule distributions. Without loss of generality, we describe our model for the distribution of rules associated with the nonterminal symbol B_j, which we denote by $\mathcal{P}_{R_j}$. The distributions over rules associated with other nonterminal symbols $B_i, i \neq j$ are modelled in an identical way.

We model the distribution $\mathcal{P}_{R_j}$ with a Dirichlet process, with base measure H_2 and with scaling parameter α_2. Since rules from $\mathcal{R}$ include elements from $\mathcal{A}, \mathcal{B}$, and $\mathcal{F}$, the distributions $\mathcal{P}_{\mathcal{A}}$, $\mathcal{P}_{\mathcal{B}}$ and $\mathcal{P}_{\mathcal{F}}$ are nested within the distribution $\mathcal{P}_{\mathcal{R}}$. Again using standard notation, we have

$$\mathcal{P}_{R_j}|H_2 \sim DP(\alpha_2, H_2)$$

$$H_2|p_e, p_\epsilon, \mathcal{P}_\mathcal{A}, \mathcal{P}_\mathcal{F} = (1-p_e)\,\mathcal{P}_\mathcal{B}^{\bigotimes 2} \otimes \mathcal{P}_\mathcal{F} + p_e\left((1-p_\epsilon)\mathcal{P}_\mathcal{A}^{\bigotimes 2} + \frac{p_\epsilon}{2}\left(\delta_\epsilon \otimes \mathcal{P}_\mathcal{A} + \mathcal{P}_\mathcal{A} \otimes \delta_\epsilon\right)\right)$$

$$\begin{aligned}
&p_e \sim Beta(a_e, b_e)\\
&p_\epsilon \sim Beta(a_\epsilon, b_\epsilon)\\
&\mathcal{P}_\mathcal{A} \sim Categorical(\mu_a^1, \ldots, \mu_a^{|\mathcal{A}|})\\
&\mathcal{P}_\mathcal{F} \sim Categorical(\mu_f^1, \ldots, \mu_f^{120})\\
&(\mu_a^1, \ldots, \mu_a^{|\mathcal{A}|}) \sim Dirichlet(\gamma_a^1, \ldots, \gamma_a^{|\mathcal{A}|})\\
&(\mu_f^1, \ldots, \mu_f^{120}) \sim Dirichlet(\gamma_f^1, \ldots, \gamma_f^{120}).
\end{aligned}$$

Since the base distribution H_2 is quite complex, let us describe it in words. Under H_2, a rule will be an emission rule (Equation 3.5) with probability p_e and a production rule (Equation 3.4) with probability $(1-p_e)$.

- If a rule is an emission rule, with probability $(1-p_\epsilon)$ it will emit two terminal symbols from the distribution $\mathcal{P}_\mathcal{A}$, with probability $\frac{p_\epsilon}{2}$ it will emit the empty symbol ϵ on the left and a terminal symbol from $\mathcal{P}_\mathcal{A}$ on the right, and with probability $\frac{p_\epsilon}{2}$ it will emit a terminal symbol from $\mathcal{P}_\mathcal{A}$ on the left and the empty symbol on the right. The symbol δ_ϵ indicates the Dirac distribution on the empty symbol, ϵ.

- If a rule is a production rule, it will produce two nonterminal symbols from $\mathcal{P}_\mathcal{B}$ and one permutation from the categorical distribution $\mathcal{P}_\mathcal{F}$.

To improve model flexibility, we put prior distributions on the model's hyperparameters $p_e, p_\epsilon, \bar{\mu}_a$ and $\bar{\mu}_f$, as demonstrated above.

We model trees through the joint distribution of the rules associated with each of their internal nodes. Since sentences are deterministic given trees, this defines a model over the sentences from the language associated with a 2-MCFG.

Remark 5. *Our model consists of mutually nested Dirichlet processes. Although separate DP distributions are used for rules associated with separate nonterminal symbols, they are all closely related due to the use of the same base distribution. Since the nonterminal component of the base distribution is itself modelled with a DP, our model can be described as an adaptation of Teh et al. (2006)'s Hierarchical Dirichlet process.*

3.4 Implementation

Let $\bar{y} = (y_1, \ldots, y_T)$ be a set of observed sentences, let l_t be the length of the t^{th} sentence and let τ_t be the tree describing its structure. Let a_t^i represent the i^{th} word of the t^{th} sentence, in such a way that $y_t = a_t^{1:l_t}$.

As described in Section 3.2.2, all internal nodes of τ_t are associated with a nonterminal symbol and a rule. Let $B_{j(x)}$ denote the nonterminal symbol associated with the node x. We assume that rules are listed in the order in which they appear in the generative process for trees, as described in Section 3.7.2. We then let $q(x)$ denote the number of rules associated with the nonterminal symbol $B_{j(x)}$ that appear before the rule associated with the node x in this list, and let $R_{j(x)}^{q(x)+1}$ denote the rule associated with the node x.

The posterior distribution we wish to estimate is then

$$\Pi\left(\tau_{1:T}|\bar{y}\right) \propto \prod_{t=1}^{T} \prod_{x \in \tau_t'} \mathcal{P}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | R_{j(x)}^{1:q(x)}, B_{1:j(x)} \right),$$

where τ_t' in the above product denotes the set of internal nodes in τ_t and where $\mathcal{P}_{R_{j(x)}}$ denotes the distribution over rules with the nonterminal symbol $B_{j(x)}$ on the left-hand side (described in Section 3.3).

To estimate this distribution over trees, we use sequential Monte Carlo (SMC) (Chopin et al. (2020)). Before describing our SMC sampling scheme in detail, we formally define complete and partial trees, extension of trees, and evaluation of trees.

Definition 5. *We define a 2-MCFG complete tree to be a tree structure where all internal nodes are associated with a nonterminal symbol and a production rule, and all leaves are associated with a nonterminal symbol and an emission rule. For any subtree of a complete tree consisting of a parent node and its direct child nodes z_1, z_2, the rule associated with x must take the form*

$$B_{j(x)} \to f_{j(x)}^{q(x)+1}[B_{j(z_1)}, B_{j(z_2)}],$$

where $B_{j(x)}$ is the nonterminal symbol associated with the node x and where $B_{j(z_i)}$ is the nonterminal symbol associated with the node z_i for $i \in \{1,2\}$.

Definition 6. *We define a 2-MCFG partial tree to be a tree structure identical to a 2-MCFG complete tree, except that leaf nodes are no longer required to be associated with emission rules.*

Remark 6. *2-MCFG partial trees may be transformed into 2-MCFG complete trees by adding nonterminal symbols and rules to their leaf nodes recursively until all leaf nodes are associated with emission rules.*

From now on, we will refer to 2-MCFG complete trees and 2-MCFG partial trees as complete trees and partial trees respectively.

Definition 7. *We define an extension of some partial tree τ to be a tree that strictly contains τ and that has additional nodes branching from one or more of the leaf nodes of τ.*

Definition 8. *An evaluated complete tree is a tree structure where each node is associated with a pair of strings of terminal symbols. The sentence that the tree represents is obtained by concatenating the pair of terminal symbols associated with the root node.*

In order to simulate a sentence from some 2-MCFG, a tree must first be generated in a top-to-bottom fashion (i.e. starting at the root node and moving down). Once the tree is complete, it is evaluated from bottom to top (i.e. starting at the leaf nodes and moving up), in order to obtain the sentence at the root node. For details on how trees are constructed and evaluated, see Subsection 3.2.2. Note that the only random part of this process is the generation of the trees. Once a tree has been generated, its evaluation is deterministic.

At each time step $t < T$ of our SMC scheme, we target some intermediate distribution over the set of trees, $\{\tau_1, \dots, \tau_t\}$ (and the rules and symbols used within them). At any node $x \in \tau_t$, we define $\tau_{t,x}$ to be the partial tree consisting of all of the nodes of τ_t added before the node x in the construction of the tree τ_t (see Section 3.7.2 for details on the order at which rules are added to a tree under our SMC scheme).

At step t, the transition kernel is

$$K\left(\tau_t | \tau_{1:t-1}, y_t, l_t\right) = \prod_{x \in \tau_t'} \mathcal{P}^{\star}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | \tau_{t,x}, R_{j(x)}^{1:q(x)}, B_{1:j(x)}, y_t, l_t \right)$$

where, as before, $j(x)$ denotes the nonterminal symbol associated with the node x, $q(x)+1$ denotes the index of the rule associated with the node x, the term $B_{1:j(x)}$ denotes the set of "observed" nonterminals, and $R_{j(x)}^{1:q(x)}$ denotes the set of "observed" rules which have $B_{j(x)}$ on the left-hand-side. The distributions $\mathcal{P}^{\star}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | \tau_{t,x}, R_{j(x)}^{1:q(x)}, B_{1:j(x)}, y_t, l_t \right)$ are described in detail in Section 3.7.3 in a case by case fashion. They are closely related to the distributions $\mathcal{P}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | R_{j(x)}^{1:q(x)}, B_{1:j(x)} \right)$ of our model, but have extra constraints due to the fact that, when evaluated, it must represent the sentence y_t.

We re-weight with

$$w_t = \prod_{x \in \tau_t'} \tilde{w}_x,$$

where $\tilde{w}_x$ is described in Section 3.7.3. By the careful design of the factors $\tilde{w}_x$ of the weights, we have

$$\mathcal{P}^{\star}_{R_{j(x)}}\left(R^{q(x)+1}_{j(x)}|\tau_{t,x},R^{1:q(x)}_{j(x)},B_{1:j(x)},y_t,l_t\right)\times\tilde{w}_x=\mathcal{P}_{R_{j(x)}}\left(R^{q(x)+1}_{j(x)}|R^{1:q(x)}_{j(x)},B_{1:j(x)}\right).$$

It follows that

$$K\left(\tau_t|\tau_{1:t-1},y_t,l_t\right)\times w_t=\prod_{x\in\tau'_t}\mathcal{P}_{R_{j(x)}}\left(R^{q(x)+1}_{j(x)}|R^{1:q(x)}_{j(x)},B_{1:j(x)}\right).$$

At iteration t, we are thus targeting

$$\begin{aligned}\Pi_t(\tau_t|y_t,l_t)&\propto\Pi_{t-1}\left(\tau_{t-1}|y_{t-1},l_{t-1}\right)\times K\left(\tau_t|\tau_{1:t-1},y_t,l_t\right)\times w_t\\&\propto\Pi_{t-1}\left(\tau_{t-1}|y_t,l_t\right)\times\prod_{x\in\tau'_t}\mathcal{P}_{R_{j(x)}}\left(R^{q(x)+1}_{j(x)}|R^{1:q(x)}_{j(x)},B_{1:j(x)}\right)\end{aligned}$$

and at the end, we reach

$$\Pi\left(\tau_{1:T}|\bar{y}\right)=\prod_{t=1}^{T}\prod_{x\in\tau'_t}\mathcal{P}_{R_{j(x)}}\left(R^{q(x)+1}_{j(x)}|R^{1:q(x)}_{j(x)},B_{1:j(x)}\right)$$

as hoped.

3.5 Simulation studies

We illustrate our model and inference method with two simulation studies: one using artificial data from the copy grammar of Example 4 and one using real data that describes the vocalizations recorded from Muriqui monkeys.

3.5.1 Artificial data: copy grammar

We consider artificial data consisting of T sentences each of length 30 simulated from the copy grammar of Example 4, i.e. from the set $\mathcal{G}_C=\{\zeta^2|\zeta\in\mathcal{A}^{\star}\}$, where $\mathcal{A}$ is defined to be an alphabet consisting of three letters: $\mathcal{A}=\{a,b,c\}$. The sentences were

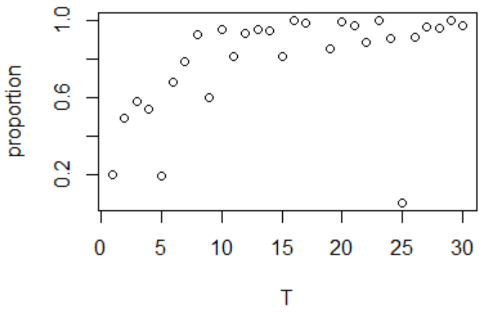


Figure 3.3: Proportion of the 1000 sentences simulated from the posterior predictive that belong to $\mathcal{G}_C$, for various values for the number of observed sentences, T ranging from 1 to 30, where posteriors were based on observed sentences of length 30 and 1000 particles.

simulated in such a way that each of the letters of the alphabet appeared with equal probability.

We applied our model and SMC algorithm to this data, using 1000 particles. We used the following Beta hyperparameters for the prior probability of an emission rule: $a_e = b_e = 10$ and the following Beta hyperparameters for the prior probability of emitting the empty symbol: $a_\epsilon = 10$ and $b_\epsilon = 1$. We used the following Dirichlet hyperparameters for the prior probability over terminal symbols: $\mu_a^i = 1, i \in \{1, \ldots, 3\}$, and the following Dirichlet hyperparameters for the prior probability over permutations: $\mu_f^i = 0.01, i \in \{1, \ldots, 120\}$.

First, with $\alpha_1 = \alpha_2 = 0.5$, we considered posterior estimation from various values of T (the number of observed sentences), ranging from 1 to 30. In each setting, 1000 sentences were simulated from the posterior predictive distribution. Figure 3.3 displays the proportion of the simulated sentences that belong to the copy grammar $\mathcal{G}_C$, for each of the settings of T. Observe that the proportion of correctly estimated sentences increases with the number of observations. This indicates that (with enough data), the posterior concentrates on grammars that represent the copy language.

Next, with $T = 200$, we considered various different settings for the Dirichlet process

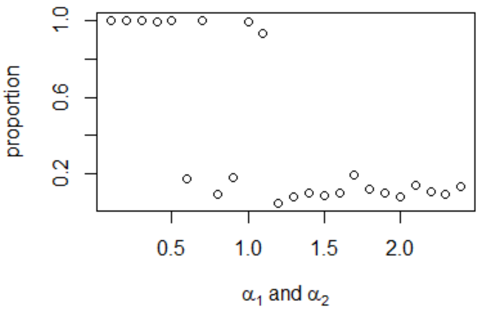


Figure 3.4: Proportion of the 1000 sentences simulated from the posterior predictive that belong to $\mathcal{G}_C$, for various values of $\alpha_1 = \alpha_2$ ranging from 0.1 to 2.4, where posteriors were based on 200 observed sentences of length 30 and 1000 particles.

parameter over rules, α_1, and the Dirichlet process parameter over nonterminal symbols, α_2, ranging from 0.1 to 2.4, where for each setting, these two parameters were equal to each other. In each setting, 1000 sentences were simulated from the posterior predictive distribution. Figure 3.4 displays the proportion of the simulated sentences that belong to the copy grammar $\mathcal{G}_C$, for each of the settings of α_1 and of α_2. Observe that for small values of α_1 and α_2 the proportion is very high, whereas as soon as α_1 and α_2 exceed 0.5, the majority of the time, the proportion is very low. This could be explained by the fact that large Dirichlet process parameters α_1 and α_2 favor more complex equivalent forms of $\mathcal{G}_C$ with more rules and nonterminal symbols. A larger number of observations is necessary to make inference on a more complex grammar.

Finally, we considered the particular setting where $T = 200$ and $\alpha_1 = \alpha_2 = 0.5$, and looked more closely at the rules of the MAP grammar. Over 99% of the emission rules were of one of the following three forms:

$$B_1 \rightarrow (c, c)$$

$$B_1 \rightarrow (b, b)$$

$$B_1 \rightarrow (a, a)$$

and over 99% of the production rules were of the form

$$B_1 \to f_1^1[B_1, B_1],$$

where f_1^1 is the permutation (13524), coinciding perfectly with the copy grammar as defined in Example 4.

3.5.2 Real data: Muriqui monkey grammar

We consider real data consisting of 647 "sentences" of Muriqui monkey vocalizations, recorded and transcribed by experts (Demolin et al. (2016)). After the consolidation of the data, the set of "words" is the alphabet of three different types of calls of the monkeys, which we denote by X, r, and p : we have $\mathcal{A} = \{X, r, p\}$.

Defining an "rp block" to be any maximal substring of the form rp^+ (i.e. one r followed by a string of p's of length at least equal to one), we say that a sequence is an *increasing rp sequence* if each rp block in it is either the same length or one p longer than the previous rp block, i.e. if the sequence is of the form

$$(rp)^{\xi_1}(rp^2)^{\xi_2}\cdots(rp^N)^{\xi_N} \ \ \xi_1, \ldots, \xi_N \geq 1, \ \ N \in \mathbb{N}.$$

Demolin et al. (2016) empirically observed the presence of such increasing rp sequences in this data, and noted that they were significantly more prevalent than decreasing rp sequences (defined similarly). Following this observation, Chatain et al. (2023) show using probabilistic methods that the language of Muriqui monkeys is likely to be more complex than the standard class of context-free grammars. In particular, they show that when sentences are simulated from the optimal context-free grammar (under the BIC criterion), the proportion of increasing sequences in the simulated data is significantly lower than the proportion of increasing sentences in the data (with empirical p-value $< 10^{-7}$) and that, furthermore the sub-language consisting of increasing sequences (without any noise) is context-sensitive, i.e. more complex the standard context-free

rrrrprprpr
$\text{rrr}\underline{\text{rprprprprprprpp}}$
$\text{rXp}\underline{\text{rprprprprprpp}}$
$\text{rXp}\underline{\text{rprprprprprpp}}\text{p}$
$\text{XXp}\underline{\text{rprprprprprpprpp}}\text{XX}$
XrpXrprpXrprprpXX
$\text{Xrpprpprpp}\underline{\text{rprpp}}\text{rpXpX}$
XXXpprpprpprpXXX
XXXpprpprpprpXXX
$\text{X}\underline{\text{rprpprpp}}\text{rp}$

rrrrprprprprp
rrprprpXpXpXrr
$\text{X}\underline{\text{rprprprprprprpp}}\text{rpXX}$
$\text{rrrp}\underline{\text{rprprpp}}\text{rp}$
$\text{Xp}\underline{\text{rpprpprppprppp}}$
$\text{Xpp}\underline{\text{rprprprprpp}}$
XXXrrpXpXprpprpprpprrXX
XXXrrpXpXprpprpprpprrXX
XXprpprppXprpprpprpX
$\text{X}\underline{\text{rprpprpprppprppp}}$

Figure 3.5: A random sample from the subset of the data with at least three consecutive rp^+ blocks. The increasing rp subsequences are underlined

class.

The goal of this simulation study is to investigate how well our model and SMC inference scheme pick up the increasing rp sequence feature of the data. Due to the limitations of computational time and memory, we consider as our observations only the subset of the data consisting of the 105 sentences that contain at least three consecutive rp^+ blocks, i.e. sentences that contain subsequences of the form $(rp^+)^3$. Figure 3.5 provides a random sample from this reduced dataset, with all increasing subsequences underlined.

We apply our 2-MCFG model and SMC inference scheme to the data, using 20000 particles and using the same model hyperparameters as in the Section 3.5.1: $a_e = b_e = 10$, $a_\epsilon = 10$, $b_\epsilon = 1$, $\mu_a^i = 1, i \in \{1, \ldots, 3\}$, $\mu_f^i = 0.01, i \in \{1, \ldots, 120\}$, and $\alpha_1 = \alpha_2 = 0.5$. We then simulate 1000 sentences from the posterior predictive distribution, and for the sake of comparison, we simulate 1000 sentences from the prior predictive distribution.

Figure 3.6 shows the proportion of sentences in the observed data, in simulated data from the prior predictive distribution, and in simulated data from the posterior predictive distribution that contain N consecutive rp^+ blocks for $N \in \{1, \ldots, 10\}$. Note that the reason why the proportion is 100% for the observed data for $N \in \{1, 2, 3\}$ is because we took, as our observations, only sentences that had at least 3 consecutive

	Data	Prior predictive	Posterior predictive
Increasing	0.65	0.05	0.21
Decreasing	0.18	0.04	0.13

Table 3.1: Proportion of sentences: in the data (left), simulated from the prior predictive (middle), and simulated from the posterior predictive (right) that contain increasing *rp* sequences (first row) and decreasing *rp* sequences (second row)

rp^+ blocks. We observe that the proportion of rp^+ blocks of every length is greater in the data simulated from the posterior predictive than in the data simulated from the prior predictive (yet still not as great as in the observed data).

We next look more specifically at the proportion of sentences that have at least one increasing *rp* sequence (of any length), and the proportion of sentences that have at least one decreasing *rp* sequence (of any length). Table 3.1 provides the proportion of sentences with increasing and decreasing *rp* sequences in the observed data, in the data simulated from the prior predictive, and in the data simulated from the posterior predictive. In the data, there is a significant asymmetry between increasing and decreasing subsequences: 65 % of sentences include an increasing subsequence, but only 18 % of the sentences include a decreasing subsequence. This matches the observations of Demolin et al. (2016) and Chatain et al. (2023). Under the prior predictive, as expected, there is no significant difference between increasing and decreasing subsequences. Both types are rare, at respectively 5% and 4%. Under the posterior predictive, we observe an asymmetry between increasing and decreasing subsequences, which occur in respectively 21% and 13% of the sentences. This difference is notable, even though it is not as large as the one observed in the original data. These observations show a similar trend to that of Figure 3.6: they indicate that the posterior predictive is significantly influenced by the data for these key statistics of interest, but that the amount of data is possibly not large enough to match the massive asymmetries observed in the real data.

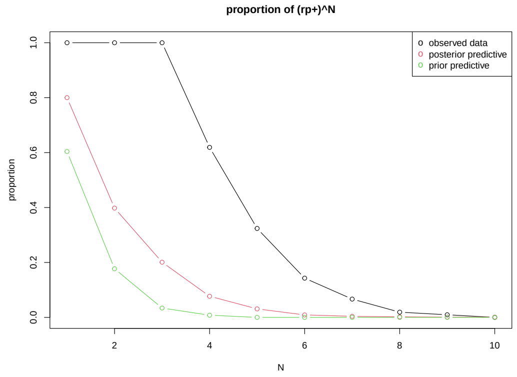


Figure 3.6: Proportion of sentences in the data (black), simulated from the prior predictive (green), and simulated from the posterior predictive (red) that contain sequences of the form $(rp^+)^N$ for $N \in \{1, \ldots, 10\}$.

3.6 Discussion

We have proposed a Bayesian model for 2-MCFGs based on the hierarchical Dirichlet process and have developed a sequential Monte Carlo algorithm to make inference. This is the first time that a Bayesian model has been applied to this class of grammars.

Our method performed well in the simulation study based on the "copy" grammar, recovering the correct grammar as soon as the Dirichlet parameters, α_1 and α_2 were reasonable (both between 0.1 and 1) and as soon as the number of observed sentences, T, was large enough (over 10).

In the case of the simulation study based on the grammar describing the linguistic structure of Muriqui monkey vocalizations, our method succeeded to some extent, in picking up the increasing sequence feature present in the language. Indeed, the proportion of sentences simulated from the posterior predictive distribution exhibiting this feature was much higher than the proportion of sentences simulated from the prior predictive distribution exhibiting this feature. However, the proportions related to the posterior predictive were significantly lower than the proportions related to the observed data, suggesting, perhaps, that 105 sentences were not enough information to fully learn every feature of the data. This is not surprising: while an exact representation of the monkey grammar in terms of 2-MCFG rules and nonterminal symbols, if it exists, remains unknown, it is unlikely that it would be as sparse as the copy grammar, which was chosen specifically for its simplicity.

It could be interesting to explore how the model or SMC algorithm could be altered to improve efficiency, and thus allow more data to be processed in a reasonable amount of time. One possible extension to our model would be to use hierarchical Pitman–Yor processes rather than hierarchical Dirichlet processes to model nonterminal symbols and rules. Indeed, numerous authors have commented on the fact that PYPs produce power-law distributions, that are closely related to those seen in natural language (Goldwater and Griffiths (2007), Teh et al. (2006)).

This work was motivated by previous results that showed in various cases that grammars describing certain natural languages are more complex than the standard class of context-free grammars. An interesting future line of researchers would be to use this model and SMC inference mechanism to calculate marginal likelihoods, and thus to make model choice using Bayes factors between context-free and 2-MCFG, in a similar way as Ryder et al. (2023) make model choice between the regular and context-free grammars.

3.7 Appendix

3.7.1 Evaluating partial trees

Recall that partial trees are trees that include one or more leaf nodes that are not associated with an emission rule. We represent each node of an evaluated partial tree by $\langle \zeta_1 | \zeta_2 \rangle$ where ζ_1 and ζ_2 are strings of terminal symbols and numbers, and where "|" is a special symbol that indicates where the string of terminal symbols and numbers $\zeta_1\zeta_2$ is split. This differs from the case of evaluated complete trees, where ζ_1 and ζ_2 must be strings of terminal symbols only.

Definition 9. *An evaluated partial tree is a tree structure where each node is associated with a pair of strings of numbers and terminal symbols. For example, some node x of an evaluated partial tree may be associated with the 2-tuple $\langle ab4, 1\rangle$, where $a, b \in \mathcal{A}$.*

An evaluated partial tree is formed from a partial tree as follows. First, whenever an emission rule $B_j \to (a_{k_1}, a_{k_2})$ takes place at some node of the partial tree, we associate that node in the evaluated partial tree with the pair $\langle a_{k_1} | a_{k_2} \rangle$. Whenever a production rule $B_j \to [B_{k_1}, B_{k_2}]$ takes place at some node, we first must define the strings $\zeta_1', \zeta_2', \zeta_3'$ and ζ_4' as follows.

- If B_{k_1} is associated with a rule, let $\langle \zeta_1 | \zeta_2 \rangle$ be the strings associated with it in the evaluated partial tree.

  – If ζ_1 is a string of terminal symbols or the empty symbol only (i.e. no numbers), set $\zeta_1' := \zeta_1$. Otherwise set $\zeta_1' := 1$.

  – If ζ_2 is a string of terminal symbols or the empty symbol only (i.e. no numbers), set $\zeta_2' := \zeta_2$. Otherwise set $\zeta_2' := 2$.

  If B_{k_1} is not associated with a rule, set $\zeta_1' := 1$ and set $\zeta_2' := 2$.

- If B_{k_2} is associated with a rule, let $\langle \zeta_3 | \zeta_4 \rangle$ be the strings associated with it in the evaluated partial tree.

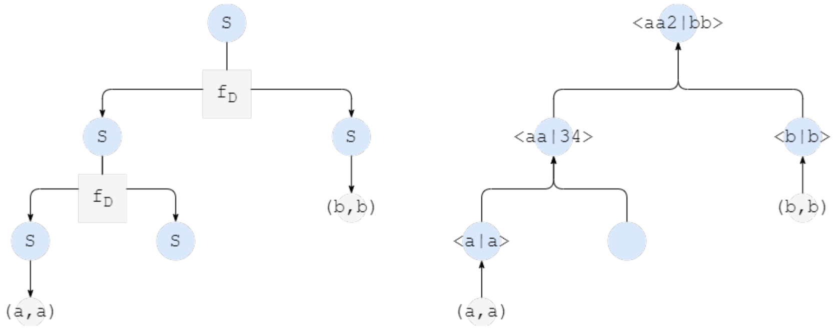


Figure 3.7: Left: A partial tree generated from the 2-multiple context-free grammar $\mathcal{G}_{D,2}$ (see Example 3). The rules used to generate the tree are (in order): $S \to f_D[S,S]$, $S \to f_D[S,S]$, $S \to (a,a)$, $S \to (b,b)$. Right: The evaluated version of the 2-MCFG partial tree.

- If ζ_3 is a string of terminal symbols or the empty symbol only (i.e. no numbers), set $\zeta_3' := \zeta_3$. Otherwise set $\zeta_3' := 3$.
- If ζ_4 is a string of terminal symbols or the empty symbol only (i.e. no numbers), set $\zeta_4' := \zeta_4$. Otherwise set $\zeta_4' := 4$.

If B_{k_2} is not associated with a rule, set $\zeta_3' := 3$ and set $\zeta_4' := 4$.

Recall that $f_j \in \mathcal{F}$ is a permutation over 5 elements. Associate B_j in the evaluated partial tree with the pair of strings $\langle f_j(\zeta_1', \zeta_2', \zeta_3', \zeta_4', |)\rangle$. For example, suppose that B_{k_1} were associated with the pair $\langle 12|a_1a_2\rangle$ (where $a_1, a_2 \in \mathcal{A}$), that B_{k_2} were not yet associated with a rule, and that $f_j = (15342)$. Then we would associate B_j with the output of the permutation f_j when applied to the vector $v := (\zeta_1', \zeta_2', \zeta_3', \zeta_4', \text{“}|\text{”}) = (1, a_1a_2, 3, 4, \text{“}|\text{”})$. In this example, the output of f_j when applied to v gives $\langle 1|34a_1a_2\rangle$.

This process is repeated from the bottom to the top of the partial tree, until every node in the evaluated partial tree is associated with a pair of strings of terminal symbols and numbers.

3.7.2 Order of the rules

Here we provide details on the order in which nodes and their associated symbols and rules are added to form the tree τ_t at time t of our SMC sampling scheme. Let's use the symbol ζ to represent some indicator (to be overwritten) that describes where we are in a partial tree at any given time. The symbol ζ will always represent both a particular node of the tree and a side (either left or right). Given ζ and an evaluated partial tree, we then can always identify a string of numbers and terminal symbols (since each node of the evaluated partial tree is associated with two strings of numbers and terminal symbols, one for each side).

We start by initialising ζ to be the left-hand side string of the root node of the partially evaluated tree. If the left-hand string of the root node is either empty or already contains terminal symbols only, we set ζ to be the right-hand string of the root node. Next, we identify the first number appearing in the string associated with ζ. For example, if the string were $cde32g$, the number to be identified would be 3. We then identify which child node and which side corresponds to the selected number. The number 3, for example, would mean to identify the left-hand side of the second child node. We reset ζ to be equal to the identified node and side. We continue moving down the tree in this way until ζ is set to be one of the strings associated with a leaf node.

We then draw a rule from the transition kernel (details are provided in Subsection 3.7.3). If this rule is a production rule, we continue as before by identifying the first integer symbol of the string associated with ζ, and overwriting ζ to be the corresponding child node and side of the extended partially evaluated tree. We extend the tree downwards like this, overwriting ζ each time as long as the rule is a production rule.

When the rule is an emission rule, we replace ζ with the emitted symbol. Note that under the transition kernel, emission rules take one of the following three forms

$$\begin{aligned}
&B_j \to (\epsilon, a_{j_2}) \\
&B_j \to (a_{j_1}, \epsilon) \\
&B_j \to (a_{j_1}, a_{j_2}),
\end{aligned}$$

where $a_{j_1}, a_{j_2} \in \mathcal{A}$. Note that since ζ only points to one side of the emission rule at a time, it is possible for one of the symbols (the symbol to which ζ points last), to initially be unspecified. It will only be when ζ points to the node and side associated with that symbol that the value it takes may be specified. When specified, it will be the word a_t^i where i is such that, at the time of emission, a_t^i has not yet been emitted, and the words $a_t^{1:i-1}$ already have been emitted.

We then re-evaluate the extended partial tree (as in Section 3.7.1). We continue the process from the beginning (starting again at the left-hand side of the root node) until all leaf nodes are associated with emission rules, and all symbols $a_t^{1:l_t}$ have been emitted. The resulting tree is τ_t is complete, and the sentence y_t may be formed by concatenating the two strings associated with its evaluated version's root node.

3.7.3 Transition kernel and weights

As described in Section 3.4, the posterior distribution we wish to estimate is

$$\Pi\left(\tau_{1:T}|\bar{y}\right) \propto \prod_{t=1}^{T}\prod_{x\in\tau_t'}\mathcal{P}_{R_{j(x)}}\left(R_{j(x)}^{q(x)+1}|R_{j(x)}^{1:q(x)},B_{1:j(x)}\right),$$

where

$$\mathcal{P}_{R_{j(x)}}\left(R_{j(x)}^{q(x)+1}|R_{j(x)}^{1:q(x)},B_{1:j(x)}\right) = \frac{1}{\alpha_2+q(x)}\left(\alpha_2 H_2 + \sum_{i=1}^{q(x)}\delta^i_{R_{j(x)}}\right)$$
$$H_2 = (1-p_e)\,\mathcal{P}_{\mathcal{B}}^{\bigotimes 2}\otimes\mathcal{P}_{\mathcal{F}} + p_e\left((1-p_\epsilon)\mathcal{P}_{\mathcal{A}}^{\bigotimes 2}+\frac{p_\epsilon}{2}\left(\delta_\epsilon\otimes\mathcal{P}_{\mathcal{A}}+\mathcal{P}_{\mathcal{A}}\otimes\delta_\epsilon\right)\right).$$

Recall that every node of an evaluated tree represents a certain number of terminal symbols in the form of two strings of terminal symbols. The partial trees $\tau_{t,x}$ for $x\in\tau_t$ have been defined such that, if they were to be extended to form a complete tree, with positive probability, the sentence associated with that complete tree would be y_t. We encode this condition by associating each node x of a partial tree with two numbers, m_x and M_x. The number m_x is the minimal number of terminal symbols that the node x of the partial tree will represent in the complete tree, and the number M_x represents the maximal number of terminal symbols that the node must represent, in order for the probability that the partial tree produces the sentence s to be nonzero. The numbers m_x and M_x depend on the current partial tree $\tau_{t,x}$ and the length of the final sentence l_t, and are updated every time when new rules are added to the partial tree. See Section 3.7.4 for details on how this is done.

Mathematically, we have the following expression for the transition kernel.

$$K(\tau_t|\tau_{1:t-1}, y_t, l_t) = \prod_{x\in\tau_t'} \mathcal{P}^{\star}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | \tau_{t,x}, R_{j(x)}^{1:q(x)}, B_{1:j(x)}, y_t, l_t \right)$$

where the probabilities in the above product may be expressed in the form

$$\mathcal{P}^{\star}_{R_{j(x)}} \left(R_{j(x)}^{q(x)+1} | \tau_{t,x}, R_{j(x)}^{1:q(x)}, B_{1:j(x)}, y_t, l_t \right) = \left(1 - p_e^{\star}\right) H_{2,p}^{\star} + p_e^{\star} H_{2,e}^{\star},$$

where $p_e^{\star}$ is a probability in $[0, 1]$ and $H_{2,p}^{\star}$ and $H_{2,e}^{\star}$ are distributions. The quantities $p_e^{\star}, H_{2,p}^{\star}$ and $H_{2,e}^{\star}$ depend on the conditions m_x and M_x and are provided explicitly at the end of this section.

Under the transition kernel $P^{\star}_{R_{j(x)}}$, certain rules in $\mathcal{R}$ are not allowed (due to the constraints as described above), and are given zero probability. Production rules that are allowed are given probability proportional to their probability under H_2. In other words, there exists some constant $\alpha^{\star} \in (0, 1]$ (provided below) such that for any production rule R^p in the support of the transition kernel, the following holds.

$$\alpha^{\star} H_{2,p}^{\star}(R^p) = \alpha_2 H_2(R^p)$$

After any production rule, the particle is re-weighted with

$$w^{\star} = \frac{\alpha^{\star}}{\alpha_2 + q(x)} \left(1 + \frac{q(x)^{\star}}{\alpha_2 H_2(R^p)} \right),$$

where $q(x)^{\star}$ is the number of times that the rule R^p has already been "observed", and $H_2(R^p)$ is the density of R^p under the base distribution H_2. The factor $\frac{q(x)^{\star}}{H_2(R^p)}$

adjusts the weight to take into account the fact that under the target distribution, extra probability is given to rules that have already been observed. The factor $\frac{1}{\alpha_2+q(x)}$ is a normalizing constant.

Because of the way the order at which the rules involved in the transition kernel are emitted is defined (see Section 3.7.2), when terminal symbols from $\mathcal{A}$ are emitted under the transition kernel, the values that they take are completely deterministic (they will be the terminal symbols that correspond to the words in the sentence in the order at which they are processed). Because of this, under the transition kernel, emission rules that are allowed are given probability proportional to a variant of H_2 that does not involve the distribution $\mathcal{P}_{\mathcal{A}}$: there exists some constant $\alpha^\star \in (0,1]$ such that for any emission rules R^e in the support of the transition kernel, the following holds

$$\begin{aligned}
&\alpha^\star H^\star_{2,e}(R^e) = \alpha_2 H_{2\setminus\mathcal{A}}(R^e)\\
&H_{2\setminus\mathcal{A}}(R^e) := (1-p_\epsilon)\delta_{\mathcal{A}}^{\bigotimes 2}(R^e) + \frac{p_\epsilon}{2}\left(\delta_\epsilon \otimes \delta_{\mathcal{A}} + \delta_{\mathcal{A}} \otimes \delta_\epsilon\right)(R^e),
\end{aligned}$$

where $\delta_{\mathcal{A}}$ is the Dirac distribution on a symbol from the set $\mathcal{A}$.

After an emission rule, in the case where one of the two emitted symbols symbols is the empty symbol ϵ, and one of the emitted symbols is an element $a_{j_1} \in \mathcal{A}$, the particle is re-weighted with

$$w^\star = \frac{\alpha^\star}{\alpha_2 + q(x)}\left(\mu_a^{j_1} + \frac{q(x)^\star}{\alpha_2 H_2(R^e)}\right),$$

and in the case where both of the emitted symbols are elements $a_{j_1}, a_{j_2} \in \mathcal{A}$, the particle is re-weighted with

$$w^\star = \frac{\alpha^\star}{\alpha_2 + q(x)}\left(\mu_a^{j_1}\mu_a^{j_2} + \frac{q(x)^\star}{\alpha_2 H_2(R^e)}\right),$$

where $\mu_a^{j_1} = \mathcal{P}_{\mathcal{A}}(a_{j_1})\mu_a^{j_2} = \mathcal{P}_{\mathcal{A}}(a_{j_2})$, and where $H_2(R^e)$ is the probability density function associated with the base distribution, H_2, evaluated at R^e. The factors $\mu_a^{j_1}$ and $\mu_a^{j_2}$ adjust the weight to take into account the fact that under the target distribution, the emitted terminal symbols are not deterministic and are distributed according to $\mathcal{P}_{\mathcal{A}}$. As in the case of production rules, the factor $\frac{q(x)^\star}{H_2(R^e)}$ adjusts the weight to take into account the fact that under the target distribution, extra probability is given to rules that have already been observed and the factor $\frac{1}{\alpha_2+q(x)}$ is a normalizing constant.

In order to write down $\alpha^\star, p_e^\star, H_{2,p}^\star$ and $H_{2,e}^\star$ in closed form, we consider separately all of the different cases, in terms of the constraints m_x and M_x at node x.

1. $m_x = 1, M_x > 2$

   The rule may be any type of emission rule or production rule.

$$
\begin{aligned}
&H_{2,p}^\star = \mathcal{P}_{\mathcal{B}}^{\bigotimes 2} \otimes \mathcal{P}_{\mathcal{F}} \\
&H_{2,e}^\star = (1-p_\epsilon)\delta_{\mathcal{A}}^{\bigotimes 2} + \frac{p_\epsilon}{2}\left(\delta_\epsilon \delta_{\mathcal{A}} + \delta_{\mathcal{A}}\delta_\epsilon\right) \\
&p_e^\star = p_e \\
&\alpha^\star = \alpha_2
\end{aligned}
$$

2. $m_x = 1, M_x = 2$

   The rule cannot be a production rule but may be any type of emission rule.

$$
\begin{aligned}
&H_{2,e}^\star = (1-p_\epsilon)\delta_{\mathcal{A}}^{\bigotimes 2} + \frac{p_\epsilon}{2}\left(\delta_\epsilon \delta_{\mathcal{A}} + \delta_{\mathcal{A}}\delta_\epsilon\right) \\
&p_e^\star = 1 \\
&\alpha^\star = \alpha_2 p_e
\end{aligned}
$$

3. $m_x = 1, M_x = 1$

   The rule must be an emission rule emitting one symbol from $\mathcal{A}$ and one empty symbol ϵ.

$$\begin{aligned}
&H^{\star}_{2,e} = \frac{1}{2}(\delta_{\epsilon}\delta_{\mathcal{A}} + \delta_{\mathcal{A}}\delta_{\epsilon}) \\
&p^{\star}_{e} = 1 \\
&\alpha^{\star} = \alpha_2 p_e p_{\epsilon}
\end{aligned}$$

4. $m_x = 2, M_x = 2$

   The rule must be an emission rule emitting two symbols from $\mathcal{A}$.

$$\begin{aligned}
&p^{\star}_{e} = 0 \\
&\alpha^{\star} = \alpha_2 p_e (1 - p_{\epsilon})
\end{aligned}$$

5. $m_x = 2, M_x > 2$

   The rule may be an emission rule emitting two symbols from $\mathcal{A}$, or may be any type of production rule.

$$\begin{aligned}
&H^{\star}_{2,p} = \mathcal{P}_{\mathcal{B}}^{\bigotimes 2} \otimes \mathcal{P}_{\mathcal{F}} \\
&H^{\star}_{2,e} = \delta_{\mathcal{A}}^{\bigotimes 2} \\
&p^{\star}_{e} = (1 - p_{\epsilon})p_e / (1 - p_e + (1 - p_{\epsilon})p_e) \\
&\alpha^{\star} = \alpha_2 (1 - p_e + (1 - p_{\epsilon})p_e)
\end{aligned}$$

6. $m_x > 2, M_x > 2$

   The rule must be a production rule.

$$\begin{aligned}
&H^{\star}_{2,p} = \mathcal{P}_{\mathcal{B}}^{\bigotimes 2} \otimes \mathcal{P}_{\mathcal{F}} \\
&p^{\star}_{e} = 0 \\
&\alpha^{\star} = \alpha_2 (1 - p_e)
\end{aligned}$$

3.7.4 Updating m_x and M_x

At the root of the tree, m_x and M_x will not change, and will both be equal to the length of the full sentence, l_t. This is because, even when the tree is not yet complete, we know that the strings of terminal symbols at the root node of the complete version of the tree together will be equal to the final sentence.

Whenever new rules are added to a partial tree, m_x and M_x of all nodes x of the partial tree must be updated. Updates first take place locally at the node at which a new rule has been applied, and then at the remaining nodes in the tree as a consequence of this change. Throughout this subsection, we use the notation z_i to denote the i^{th} child of some node (for $i \in \{1, 2\}$).

First, let's look at the updates that take place locally when a rule is applied at some node of the tree. Recall that there are two different types of rules possible in our sampling scheme: production rules, and emission rules. We consider these two types separately.

1. Production rule at the node x:

   The value m_x is updated as follows.

   - $m_x \leftarrow \max(m_x, 2)$

   The maxima and minima of the child nodes of the new production rule are then initialised as follows.

   - $m_{z_i} \leftarrow 1, \qquad i \in \{1, 2\}$
   - $M_{z_i} \leftarrow M_x - 1, \qquad i \in \{1, 2\}$

2. Emission rule at the node x:

   - If the emission rule consists of two symbols from the set $\mathcal{A}$:

The values m_x and M_x are updated as follows.

$$M_x \leftarrow 2$$
$$m_x \leftarrow 2$$

- Else if the emission rule consists of one symbol from the set $\mathcal{A}$ and one empty symbol:

  The value M_x is updated as follows.

$$M_x \leftarrow 1$$

Now, we look at the updates that take place in the tree as a consequence of an update to a particular node. Our updates are based on the following definition.

Definition 10. *We say that a partial tree is fully updated if, for all subtrees of the partial tree consisting of one single parent x and its direct child nodes ($z_i, i \in \{1, 2\}$), the following are satisfied.*

$$m_x \geq m_{z_1} + m_{z_2} \tag{3.6}$$
$$m_{z_1} \geq m_x - M_{z_2} \tag{3.7}$$
$$m_{z_2} \geq m_x - M_{z_1} \tag{3.8}$$

and similarly

$$M_x \leq M_{z_1} + M_{z_2}$$
$$M_{z_1} \leq M_x - m_{z_2}$$
$$M_{z_2} \leq M_x - m_{z_1}$$

We demonstrate how updates are made for all nodes of the tree in the case where m_x has been updated (increased). The case where M_x is updated (decreased) is analogous.

Suppose that m_x has been increased. Updating the tree involves simply applying the following two steps.

1. - Set v be the parent node of x. Let $z_i, i \in \{1, 2\}$ be the child nodes of v (in such a way that x will be one of the $z_i, i \in \{1, 2\}$).
   - Update m_v.

     $$m_v \leftarrow \max\left(m_v, m_{z_1} + m_{z_2}\right)$$

   - Moving up the tree, reset v to be the parent node of v of the previous step, and reset z_1 and z_2 to be the child nodes of the new v.
   - Repeat the above two points until v is set to S.
2. For all subtrees of the partial tree consisting of a single parent v and its direct children $z_i, i \in \{1, 2\}$, starting from the topmost subtree and moving down the partial tree, do the following.

   $$M_{z_1} \leftarrow M_v - m_{z_2}$$
   $$M_{z_2} \leftarrow M_v - m_{z_1}$$

Step 1 ensures that Equation 3.6 of the definition of a fully updated tree is satisfied, and Step 2 ensures that Equation 3.7 and Equation 3.8 of the definition is satisfied.

Chapter 4

On consistency issues for Bayesian clustering using nonparametric priors

Abstract

Bayesian nonparametric (BNP) mixture models are popular for modelling complex data. Their posterior distributions exhibit nice theoretical properties, concentrating at the optimal minimax rate to the true data-generating distribution, and extensive research has been devoted to developing this theory. However it has been shown that these models are inconsistent for the number of clusters. In the case of Dirichlet process (DP) mixture models, this problem can be mitigated when a prior is put on the model's concentration hyperparameter α, as is common practice. We prove that Pitman–Yor process (PYP) mixture models (which generalise DP mixture models) remain inconsistent for the number of clusters when a prior is put on α, in the special case where the true number of components in the data generating mechanism is equal to 1 and the discount parameter $\sigma \in (0, 1)$ is a fixed constant.

When considering the space over partitions induced by BNP mixture models, point estimators such as the maximum a posteriori (MAP) are commonly used to summarise the posterior clustering structure of such models, which alone can be complex and difficult to interpret. We prove consistency of the MAP partition for DP mixture models when the concentration parameter, α_n, goes deterministically to zero, and when the true partition is made of only one cluster.

4.1 Introduction

Mixture models, popular for their flexibility and simplicity, are commonly used in the statistical analysis of heterogeneous data where observations are assumed to come from a number of different populations. Since in a mixture, each observation is assumed to come from one population, such models naturally induce a clustering: two data points belong to the same cluster if they come from the same population.

Classical methods for cluster analysis include agglomerative hierarchical clustering (where two groups chosen to optimize some criterion are merged at each stage of the algorithm) or K-means clustering (where data points are moved from one group to another until there is no further improvement in the sum of squares criterion, MacQueen (1967)). Although there has been a considerable amount of research into the development of these classical methods, they are largely heuristic.

Another solution is to use model-based methods, for which statistical properties may be formally inferred. In finite mixture models, the clusters are related to the components of mixtures. Fraley and Raftery (2002) review a general methodology for model-based clustering using finite mixture models. Data is fit with a number of different mixture models, each with a different number of components, and the best model is selected in terms of some criterion, such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC). This method, however, may be computationally expensive since many models must be fit. A Bayesian approach could alternatively be taken by putting a parametric prior (such as a Poisson) on the number of components, but inference can be challenging when the dimensionality or the amount of data becomes large (although new strategies have been proposed recently, see Miller and Harrison, 2018).

In this work, we consider infinite mixture models where the mixing measure is modeled with a nonparametric prior. In such models, the number of components possible has no upper bound. Inference may be performed in a unified way without the need for strong assumptions on the number of components and with no need to fit multiple models.

While the most standard nonparametric prior remains the Dirichlet process (DP) introduced by Ferguson (1973), many extensions now exist. The Pitman–Yor process (PYP, Pitman and Yor, 1997) is a natural extension of the DP with an extra parameter increasing model flexibility. Compared with DP mixtures, PYP mixtures are better suited when the sizes of clusters are more evenly distributed. Due to the interpretability of their hyperparameters, ease of implementation, and nice mathematical properties, Bayesian nonparametric (BNP) priors are widely used in practice, and in the last two decades, a huge amount of research has focused on their properties (see for example Ghosal and van der Vaart, 2017; Müller et al., 2018). The use of the DP as a mixing measure was first introduced by Lo (1984). Thanks to the wide variety of efficient computational methods which have been introduced for their inference (Escobar and West, 1998; MacEachern and Müller, 1998; Neal, 2000; Blei et al., 2006), nonparametric mixture models have become common in a wide range of modelling applications.

In the context of density estimation, under certain conditions the posterior distribution of DP mixture models concentrates at the true data-generating density at the minimax-optimal rate in L1 and Hellinger norms (Ghosal and van der Vaart, 2017; Ghosal et al., 1999). This holds for other types of Bayesian nonparametric priors, such as PYP priors (Lijoi et al., 2005). Nguyen (2013) further proved posterior consistency of the mixing distribution in the Wasserstein metric for DP and PYP mixture models.

It is important to realise that consistency of the posterior distribution for the data-generating density and even for the mixing measure does not imply consistency of the inferred number of clusters. Empirically, many researchers have observed that DP mixture posteriors tend to overestimate the number of clusters (West and Escobar, 1993; Lartillot and Philippe, 2004; Onogi et al., 2011). More recently, Miller and Harrison (2013, 2014) proved that the posterior distribution on the number of clusters does not concentrate to the number of components in DP and PYP mixtures. Alamichel et al. (2022) extended this result to the case of Gibbs-type processes. A possible explanation for this inconsistency result can be found in a result proved by Rousseau and Mengersen (2011), that in overfitted finite or infinite mixture models, the weights attributed to extra clusters go to zero as the number of observations grows. Provided that the weights for the extra components are infinitesimally small, any mixture can be approximated arbitrarily well by a mixture with a larger number of components.

Despite the above inconsistency results, it is possible to achieve posterior consistency for the number of clusters in the mixture models we consider. Guha et al. (2021) introduce a fast and simple post-processing procedure for DP mixtures which provides clustering consistency. Alamichel et al. (2022) extend this result to PYP mixtures. Ascolani et al. (2022) show that posterior consistency for the number of clusters can be achieved in certain cases for a DP mixture model by putting a prior on the DP concentration parameter α. DP mixtures modeled in this way can be considered as mixtures of DP mixtures (Antoniak, 1974) and are commonly used in practice.

Beyond the distribution over the number of clusters, an interesting question in cluster analysis is the distribution over the partition space across clusters induced by BNP mixture models. This space is large and complex: the number of possible clusterings of n items grows exponentially according to $B(n)$, the Bell number of n items (Bell (1934)). Since it would be infeasible to describe the posterior density of all the unique partitions, it is common practice to find a point estimator to concisely represent the posterior.

Some authors have proposed BNP model clustering estimators based on pairwise probabilities that items belong to the same cluster. Medvedovic et al. (2001) and Medvedovic and Sivaganesan (2002) have proposed methods that make use of the posterior similarity matrix. For a sample of size n, the elements of this $n \times n$ matrix represent the probability that two data points are in the same cluster. Classical hierarchical clustering algorithms are then applied based on this similarity matrix. The disadvantage of these methods is that they require sampling from the posterior clustering distribution (usually through Markov chain Monte Carlo). Posterior probabilities of individual partitions are difficult to compute reliably from Monte Carlo samples.

Another approach is to find the partition that minimizes the posterior risk associated to some loss function. That is, the partition that minimizes

$$l(c'|y) = \sum_{c \in \mathcal{C}} l(c', c)\pi(c|y)$$

for some choice of loss function $l(c', c)$, where $\mathcal{C}$ is the set of all possible partitions. Binder (1978) discusses loss functions for Bayesian clustering. Dahl (2006) and Lau and Green (2007) propose loss functions that penalize miss-assigned groups. Wade and Ghahramani (2018) provide a discussion on types of loss functions and methods of finding the optimal partition.

The optimal Bayes estimate of the clustering under the 0-1 loss function is equivalent to the maximum a-posteriori (MAP) clustering estimator (Binder (1978)), and is commonly used in Bayesian model-based procedures (Broët et al. (2002), Kim et al. (2006), Li et al. (2007)). The 0-1 loss function may be described intuitively as follows: no loss is incurred if the clustering estimate equals the true clustering and a loss of one is incurred for any other clustering estimate. Many fast algorithms have been developed for finding the MAP estimator, making it often more convenient than other estimator choices in practice. Dahl (2009) proposes a fast and efficient search algorithm that is guaranteed to find the MAP clustering for univariate product partition models (of which the Dirichlet process mixture model is a special case when one integrates over the model parameters). Fuentes-García et al. (2019) propose an alternative algorithm for finding the MAP clustering which is motivated by Hopfield's network.

Rajkowski (2019) investigate theoretical properties of the MAP partition in the particular case of Gaussian Dirichlet process mixture models (where the cluster means have Gaussian distribution and, for each cluster, the observations within the cluster have Gaussian distribution). Along with some nice theoretical properties, they prove that model mis-specification can lead to non-consistency of the MAP partition.

Our contributions are as follows. In Theorem 4.3.0.1 we show that Ascolani et al. (2022)'s result cannot be directly extended to PYP mixtures: we prove inconsistency for the number of clusters for well-specified Pitman–Yor process mixture models with a prior on the concentration parameter α, when the true number of clusters in the data generating mechanism, t, is equal to one, and when the discount parameter $\sigma \in (0, 1)$ is a fixed constant. In Theorem 4.3.0.2, we prove consistency of the MAP partition for well-specified DP mixture models when the concentration parameter α_n goes to zero at an appropriate rate, and when the true partition is made of only one cluster.

Note that these two results deal with quite different settings.

- (i) While Theorem 4.3.0.1 considers the distribution over the *number of clusters* induced, Theorem 4.3.0.2 considers the *MAP estimator* over the whole space of *partitions*.

- (ii) While Theorem 4.3.0.1 deals with PY mixture models, Theorem 4.3.0.2 deals with DP mixture models.

- (iii) While in Theorem 4.3.0.1 we assume a prior distribution on the PY concentration parameter α, in Theorem 4.3.0.2 the DP concentration parameter α_n goes to zero as $n \to \infty$.

The remainder of this chapter is organized as follows. In Section 4.2 we recall formal definitions for Dirichlet process and Pitman–Yor process mixture models, and introduce the notation that we will be using throughout the chapter. In Section 4.3 we present our two theoretical results, whose proofs are given in Section 4.4. In Section 4.5 we provide a short discussion. The proofs of the lemmas used in our proofs, as well as a short simulation study to illustrate Theorem 4.3.0.1, are left to the appendix sections 4.6.1 and 4.6.2 respectively.

4.2 Preliminaries

Formally, we assume data $y_{1:n}$ is iid from a distribution $P^\star$ with pdf with respect to some measure μ

$$p^\star(y) = \sum_{j=1}^{t} \rho_j^\star k(y|\phi_j^\star) \qquad\qquad t \in \mathbb{N}, \tag{4.1}$$

where the $\rho_j^\star$ are probability weights in $(0,1)$ summing to one, and where the $k(\cdot|\phi_j^\star)$ are probability kernels, each depending on some parameter $\phi_j^\star$. The above may alternatively be expressed as a convolution of the component-specific kernel $k(\cdot|\phi)$ with the discrete mixing measure $G^\star = \sum_{j=1}^{t} \rho_j^\star \delta_{\phi_j^\star}$:

$$p^\star(y) = \int k(y|\phi) G^\star(\mathrm{d}\phi).$$

We consider the well-specified case where the kernel density $k(\cdot|\phi)$ is known, but where the integer t, the weights $\rho_j^\star$, and the latent variables $\phi_j^\star$ in Equation (4.1) are all unknown. In order to allow for an unbounded number of components t in the mixture, we consider nonparametric mixture models, with nonparametric priors on the mixing measure G.

The most standard BNP prior is the Dirichlet process (DP). When G is a draw from a DP, we write $G \sim DP(\alpha, Q_0)$, where $\alpha > 0$ is the concentration parameter and where Q_0 is the base distribution. The DP is characterized by the generative distribution of data points drawn from it: if $(\phi_1, \dots, \phi_n, \phi_{n+1}) \sim G$ and $G \sim DP(\alpha, Q_0)$, then conditional on $(\phi_1, \dots, \phi_n)$, the $(n+1)^{th}$ observation ϕ_{n+1} is equal to ϕ_j with probability $\frac{n_j}{\alpha+n}$ (where n_j represents the number of components in $(\phi_1, \dots, \phi_n)$ that take the same value as ϕ_j) and is distributed according to Q_0 with probability $\frac{\alpha}{\alpha+n}$.

The Pitman–Yor process (PYP) is a generalization of the DP, with an extra parameter that allows for the sizes of clusters to be more evenly distributed. When G is a draw from a PYP, we write $G \sim PYP(\alpha, \sigma, Q_0)$, where α and Q_0 are as in for the case of DPs and where σ is the additional discount parameter. The PYP is also characterized by the generative distribution of data points drawn from it: if $(\phi_1, \dots, \phi_n, \phi_{n+1}) \sim G$ and $G \sim PYP(Q_0, \alpha, \sigma)$, then conditional on $(\phi_1, \dots, \phi_n)$, the $(n+1)^{th}$ observation ϕ_{n+1} is equal to ϕ_i with probability $\frac{n_j - \sigma}{\alpha+n}$ and is distributed according to Q_0 with probability $\frac{\alpha+n\sigma}{\alpha+n}$. If $\sigma = 0$ we get the DP.

We consider the distributions over partitions, τ induced by the DP and the PYP. For every pair of numbers $(n, s) \in \mathbb{N}^2$ with $s \leq n$, we let $\tau_s(n)$ denote the set of partitions of $\{1, \dots, n\}$ into s non empty subsets. Conditional on parameter α (and possibly of σ) DP and PYP mixture models induce the following prior distributions on the space of partition for any $n \in \mathbb{N}$, and any $A^{(n)} = \{A_1^{(n)}, \dots, A_s^{(n)}\} \in \tau_s(n), s \leq n$:

$$\Pi_{DP}(A^{(n)}|\alpha) = \frac{\alpha^s}{\alpha_{(n)}} \prod_{j=1}^{s} (a_j - 1)!, \tag{4.2}$$

$$\Pi_{PYP}(A^{(n)}|\alpha, \sigma) = \frac{\sigma^{s-1}(1 + \frac{\alpha}{\sigma})_{(s-1)}}{(1+\alpha)_{(n-1)}} \prod_{j=1}^{s} (1-\sigma)_{(a_j - 1)}, \tag{4.3}$$

respectively, where $\alpha_{(n)} = \alpha \cdots (\alpha+n-1)$ is the ascending factorial (with the convention that $\alpha_{(0)} = 1$) and $a_j = |A_j^{(n)}|$ stands for the cardinality of the set $A_j^{(n)}$. We consider partitions $A^{(n)}$ such that (1) $a_j \neq 0$ for $j = 1, \ldots, s$ (2) $A_i^{(n)} \cap A_j^{(n)} = \emptyset$ for $i \neq j$ and (3) $\cup_{j=1}^{s} A_j^{(n)} = \{1, \ldots, n\}$. The $A_j^{(n)}, j = 1, \ldots s$ represent the clusters.

Conditionally on the partition $A^{(n)}$, the probability densities of the data $y_{1:n} = (y_1, \ldots, y_n)$ and of the cluster-specific parameters $\phi_{1:s} = (\phi_1, \ldots, \phi_s)$ are

$$p(y_{1:n}|\phi_{1:s}, A^{(n)}) = \prod_{j=1}^{s} \prod_{i \in A_j^{(n)}} k(y_i|\phi_j), \quad \pi(\phi_{1:s}|A^{(n)}) = \prod_{j=1}^{s} q_0(\phi_j),$$

where q_0 is the density of the base measure Q_0 of the DP and the PYP.

As previously mentioned, BNP mixture models with fixed hyperparameters are inconsistent in the number of clusters induced. In order to achieve consistency for the number of clusters induced, Ascolani et al. (2022) consider Dirichlet process mixture models with a prior on the concentration parameter α:

$$Y_i|\phi_i \overset{\text{ind}}{\sim} k(\cdot|\phi_i), \quad \phi_i|G \overset{\text{iid}}{\sim} G \quad G|\alpha \sim \text{DP}(\alpha, Q_0), \quad \alpha \sim \zeta, \tag{4.4}$$

where ζ is a prior distribution on α.

In our Theorem 4.3.0.1, we consider an extension of Ascolani et al. (2022)'s model, which are Pitman–Yor mixture models with a prior on the concentration parameter $\alpha > 0$ and with a fixed discount parameter $\sigma \in (0, 1)$:

$$Y_i|\phi_i \overset{\text{ind}}{\sim} k(\cdot|\phi_i), \quad \phi_i|G \overset{\text{iid}}{\sim} G \quad G|\alpha, \sigma \sim \text{PYP}(\alpha, \sigma, Q_0), \quad \alpha \sim \zeta. \tag{4.5}$$

We use the standard notation K_n to denote the number of clusters in a sample of size n. Under our model (4.5), K_n has the following prior distribution

$$\Pi_{PYP}(K_n = s|\sigma) = \int \sum_{A^{(n)} \in \tau_s(n)} \Pi_{PYP}(A^{(n)}|\alpha, \sigma)\zeta(\mathrm{d}\alpha)$$

where $\Pi_{PYP}(A^{(n)}|\alpha, \sigma)$ is as Equation (4.3) above.

After having observed data $y_{1:n}$, we consider the posterior distribution $\Pi_{PYP}(K_n =$

$s|y_{1:n}, \sigma)$. To prove our result, we start with the joint distribution which, for every $s \in \mathbb{N}$ has density with respect to the product measure $\mu^{\bigotimes n}$ times the counting measure

$$p_{PYP}(y_{1:n}, s|\sigma) = \sum_{A^{(n)} \in \tau_s(n)} \Pi_{PYP}(A^{(n)}|\sigma) \prod_{j=1}^{s} p(y_{A_j^{(n)}}) \tag{4.6}$$

where $\Pi_{PYP}(A^{(n)}|\sigma) = \int \Pi_{PYP}(A^{(n)}|\alpha, \sigma)\zeta(\mathrm{d}\alpha)$ and $p(y_{A_j^{(n)}}) = \int \prod_{i \in A_j^{(n)}} k(y_i|\phi)q_0(\phi)\mathrm{d}\phi$ is the marginal likelihood for the subset of observations identified by $A_j^{(n)}$, given that they are clustered together.

Throughout, for x_n and z_n two $n-$dependent random variables such that $z_n \neq 0 \ \forall n$, we use the notation $x_n = o(z_n)$ if $\frac{x_n}{z_n} \to 0$ as $n \to \infty$.

4.3 Theoretical results

Both of our results rely on the following assumption on the base measure Q_0 of the BNP random measure.

Assumption 6. *The base measure Q_0 is absolutely continuous with respect to the Lebesgue measure, and its density q_0 is bounded.*

Throughout, we assume kernels of the form

$$k(y|\phi) = g(y - \phi), \quad y \in \mathbb{R}.$$

Our Theorem 4.3.0.1, relies on the following assumptions on the function g.

Assumption 7. *The function g is positive on some interval $[a, b]$ and 0 elsewhere.*

Assumption 8. *The function g is differentiable with bounded derivative in (a, b).*

The above two assumptions require that the kernel is a location-family distribution with positive density on a bounded support. This class is fairly general and includes as special cases uniform distributions and truncated normal distributions. Our Theorem 4.3.0.2 assumes that the kernel must either be a Gaussian, a truncated Gaussian, a uniform, or a triangular distribution, which are all common classes.

For our Theorem 4.3.0.1, we will require three additional assumptions on the prior ζ of α, identical to the assumptions used by Ascolani et al. (2022), which we re-state below.

Assumption 9. *The prior ζ is absolutely continuous with respect to the Lebesgue measure. Its density is also denoted by ζ.*

Assumption 10. *There exist ϵ, δ, β such that, for all $\alpha \in (0, \epsilon)$ it holds that $\delta\alpha^\beta \leq \zeta(\alpha) \leq \frac{\alpha^\beta}{\delta}$.*

Assumption 11. *There exist $D, \nu, \kappa > 0$ such that $\int \alpha^s \zeta(\alpha) d\alpha < D\kappa^{-s}\Gamma(\nu + s + 1)$ for every $s \geq 1$.*

As demonstrated in Ascolani et al. (2022), Assumptions (9) - (11) are satisfied by common families of distributions, such as uniform distributions over $(0, c)$ with $c > 0$, or gamma distributions with shape ν and rate κ.

Theorem 4.3.0.1. *Suppose that the kernel k, the density q_0 and the prior ζ over the concentration parameter α satisfy Assumptions (6) - (11) stated above. For every distribution on data $P^\star$ with density $p^\star$ as in (4.1), for the number of components $t = 1$, we have*

$$\Pi_{PYP}(K_n = 1 | y_{1:n}) \not\to 1$$

uniformly in $P^\star$ as $n \to \infty$.

Theorem 4.3.0.1 shows that, unlike for the case of DP mixture models, PYP mixture models with fixed nonzero σ remain inconsistent for the number of clusters even when a prior is put on the concentration parameter α. Our negative result holds for data with one mixture component ($t = 1$ when the mixture is described by (4.1)), and when the discount parameter σ is a nonzero constant in $(0, 1)$. The proof rests on analysing the ratio $\frac{\Pi(K_n=s|y_{1:n})}{\Pi(K_n=1|y_{1:n})}$, as consistency cannot hold if it does not converge to 0 as $n \to \infty$. Following the strategy of Ascolani et al. (2022), this ratio can be split into the product of two quantities, one capturing the impact of the prior distribution on the concentration parameter α, and the other independent of the prior on α. In the Dirichlet process case with a prior on α, the first quantity goes to 0 and the second remains bounded. We show that in the Pitman–Yor case, the σ parameter enters the first quantity and prevents it from vanishing as $n \to \infty$, destroying consistency and highlighting a fundamental difference between the DP and PY processes.

Theorem 4.3.0.2. *Suppose that the density q_0 satisfies Assumption (6) stated above, and suppose that the DP concentration parameter $\alpha_n = o\left(\frac{1}{\log(n)}\right)$. For every distribution on data $P^\star$ with density $p^\star$ as in (4.1), for Gaussian, truncated Gaussian, uniform, or triangular kernels k, and for the number of components $t = 1$, we have*

$$P^\star \left\{ \exists A^{(n)} \neq A^{(n)\star} \text{ such that } \frac{\Pi_{DP}\left(A^{(n)}|y_{1:n}\right)}{\Pi_{DP}\left(A^{(n)\star}|y_{1:n}\right)} > 1 \right\} \to 0$$

as $n \to \infty$ where $A^{(n)\star}$ represents the true partition of the data (i.e. one single cluster).

Theorem 4.3.0.2 considers the whole space of partitions (not just the number of clusters induced), and more specifically the MAP point estimator of that space. Our result holds when the concentration parameter, α_n, is sent deterministically to 0 at rate $\frac{1}{\log(n)}$, as is done, for example in Ohn and Lin (2023). Our result holds for Gaussian, truncated Gaussian, uniform or triangular kernels, which are all common classes.

Our result holds for data with one mixture component ($t = 1$ when the mixture is described by (4.1)), but seems to generalise to any value, t under some assumptions of the separability of the cluster components. This is a topic of ongoing work.

4.4 Proof of Theorem 4.3.0.1 and Theorem 4.3.0.2

Without loss of generality, through a linear rescaling, we assume $[a, b] = [-c, c]$. For convenience, we rewrite the assumptions on g and Q_0 as

$T1.$ $\exists m, M$ such that $0 < m \leq g(y) \leq M < \infty$ for every $y \in [-c, c]$;

$T2.$ g is differentiable on $(-c, c)$ and $\exists R$ such that $|\frac{g'(y)}{g(y)}| \leq R < \infty$ for every $y \in (-c, c)$;

$T3.$ $\exists U > 0$ such that $h(y) := q_0(y) + q_0(-y) \leq U$ for every $y \in [0, 2c]$;

$T4.$ $\exists L > 0$ such that $q_0(\phi) \geq L$ for every ϕ in a neighbourhood of $\phi_j^\star$, for every j.

4.4.1 Statement of lemmas

The proof of Theorem 4.3.0.1 relies on the following simple lemma, used by and proved by Ascolani et al. (2022). It justifies working with ratios, which allows us to avoid calculations of marginal likelihoods of the observed data.

Lemma 4.4.1.1. *The convergence $\Pi(K_n = t|y_{1:n}) \to 1$ as $n \to \infty$ holds if and only if one has*

$$\sum_{s \neq t} \frac{\Pi(K_n = s|y_{1:n})}{\Pi(K_n = t|y_{1:n})} \to 0 \quad \textit{as } n \to \infty.$$

The proof of Theorem 4.3.0.2 relies on the remaining five lemmas, stated below. The proofs for Lemma 4.4.1.2 and Lemma 4.4.1.3 are provided in the appendix. Lemma 4.4.1.4 and Lemma 4.4.1.5 are used by and proved by Ascolani et al. (2022), and we refer the reader to that paper for their proofs. Lemma 4.4.1.6 comes from Rajkowski (2019) (for the case of Gaussian kernels). We refer the reader to that paper for its proof, which is straightforward to extend to other kernels, such as truncated Gaussian, uniform or triangular kernels.

Lemma 4.4.1.2. *Let s and n be any positive integers with $s \leq n$, let x be any positive real number, and let $\{a_j\}_{1 \leq j \leq s}$ be any set of s strictly positive integers satisfying $\sum_{j=1}^{s} a_j = n$. Then,*

$$\arg \max_{\{a_j\}_{1 \leq j \leq s}} \prod_{j=1}^{s} (a_j - 1)! = \{\tilde{a}_j\}_{1 \leq j \leq s}, \tag{4.7}$$

where $\tilde{a}_1 = \tilde{a}_2 = \cdots = \tilde{a}_{s-1} = 1$, and $\tilde{a}_s = n - s + 1$ (up to a permutation). We have

$$\prod_{j=1}^{s} (\tilde{a}_j - 1)! = (n - s)!.$$

Lemma 4.4.1.3. *Let s and n be any positive integers with $s \leq n$, and let $\{a_j\}_{1 \leq j \leq s}$ be any set of s strictly positive integers satisfying $\sum_{j=1}^{s} a_j = n$. Then,*

$$\arg \min_{\{a_j\}_{1 \leq j \leq s}} \prod_{j=1}^{s} (a_j + 1) = \{\tilde{a}_j\}_{1 \leq j \leq s}, \tag{4.8}$$

where $\tilde{a}_1 = \tilde{a}_2 = \cdots = \tilde{a}_{s-1} = 1$, and $\tilde{a}_s = n - s + 1$ (up to a permutation). We have

$$\prod_{j=1}^{s} (\tilde{a}_j + 1) = 2^{s-1}(n - s + 2).$$

Lemma 4.4.1.4. *Under model (4.1) with $t = 1$, there exists $W > 0$ and $n' \in \mathbb{N}$ such that for all $n \geq n'$ it holds*

$$\int_{\mathbb{R}} \prod_{i \in \{1,\ldots,n\}} \frac{g(y_i - \phi)}{g(y_i - \phi_1^\star)} q_0(\phi) d\phi \geq \frac{W Z_n}{n\sqrt{\log(n)}}$$

where, with $y_{(n)}$ denoting the maximum observed data-point, Z_n is defined as

$$Z_n := \min[1, n\sqrt{\log(n)}\{c + \phi_1^\star - y_{(n)}\}].$$

Furthermore, $Z_n \to 1$ in $P^\star$-probability as $n \to \infty$.

Lemma 4.4.1.5. *Under $y_{1:n}$ iid from a distribution with pdf as in Equation 4.1 with $t = 1$, it holds*

$$E\left\{\prod_{j=1}^{s}\int_{\mathbb{R}^s}\prod_{i\in A_j^{(n)}}\frac{g(y_i-\phi_j)}{g(y_i-\phi_1^\star)}q_0(\phi_s)d\phi_s\right\} \leq \left(\frac{U}{m}\right)^s\prod_{j=1}^{s}\frac{1}{a_j+1}$$

where $(\phi_1, \dots, \phi_s) \in \mathbb{R}^s$, where $A^{(n)} = \{A_1^{(n)}, \dots, A_n^{(n)}\}$ is a partition of the set $\{1, \dots, n\}$ into $s \leq n$ clusters, and where m and U are defined in $T1$ and $T3$.

Lemma 4.4.1.6. *When modeling DP mixtures of the form of Equation (4.1), with Gaussian, truncated Gaussian, uniform of triangular kernels, the MAP partition divides the data into clusters whose convex hulls are disjoint. In particular, the MAP partition respects the ordering of the data.*

4.4.2 Proof of Theorem 4.3.0.1

Proof. By Lemma 4.4.1.1, it will be sufficient to prove that $\frac{\Pi(K_n=s|y_{1:n})}{\Pi(K_n=1|y_{1:n})} \not\to 0$ pointwise in $y_{1:n}$ as $n \to \infty$, for some $s > 1$. We will prove this using $s = 2$.

In order to prove our result, we make use of results of the asymptotic behavior of certain quantities under the Dirichlet process mixture model of Ascolani et al. (2022), which can be described by Equation (4.4). Throughout this proof we will thus use the subscript DP to indicate that a quantity is related to the Dirichlet process model, and we will use the subscript PYP to indicate that a quantity is related to the Pitman–Yor model, whenever there is ambiguity.

Under our Pitman–Yor mixture model, by applying Equation (4.6), we have

$$\begin{aligned}\frac{\Pi_{\text{PYP}}(K_n=2|y_{1:n})}{\Pi_{\text{PYP}}(K_n=1|y_{1:n})} &= \frac{\int \sigma\left(1+\frac{\alpha}{\sigma}\right)\frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}{\int\frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}\frac{\sum_{A^{(n)}\in\tau_2(n)}\prod_{j=1}^{2}(1-\sigma)_{(a_j-1)}p(y_{A_j^{(n)}})}{(1-\sigma)_{(n-1)}p(y_{1:n})}\\ &= C_{\text{PYP}}(n,1,2)R_{\text{PYP}}(n,1,2)\end{aligned}$$

where

$$C_{\mathrm{PYP}}(n,1,2) := \frac{\int \sigma\left(1+\frac{\alpha}{\sigma}\right)\frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}{\int \frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}$$

and

$$R_{\mathrm{PYP}}(n,1,2) := \frac{\sum_{A^{(n)}\in\tau_2(n)}\prod_{j=1}^{2}(1-\sigma)_{(a_j-1)}p(y_{A_j^{(n)}})}{(1-\sigma)_{(n-1)}p(y_{1:n})}.$$

Similarly, under the Dirichlet process mixture model of Ascolani et al. (2022), one gets

$$\frac{\Pi(K_n = s|y_{1:n})_{\mathrm{DP}}}{\Pi(K_n = t|y_{1:n})_{\mathrm{DP}}} = C_{\mathrm{DP}}(n,t,s)R_{\mathrm{DP}}(n,s,t),$$

where $C_{\mathrm{DP}}(n,t,s)$ is an integral in α over all of the terms involving α, and $R_{\mathrm{DP}}(n,t,s)$ contains all of the remaining factors:

$$C_{\mathrm{DP}}(n,t,s) := \frac{\int[\alpha^s\zeta(\alpha)/\alpha_{(n)}]d\alpha}{\int[\alpha^t\zeta(\alpha)/\alpha_{(n)}]d\alpha}$$

and

$$R_{\mathrm{DP}}(n,t,s) := \frac{\sum_{A^{(n)}\in\tau_s(n)}\prod_{j=1}^{s}(a_j-1)!\prod_{j=1}^{s}p(y_{A_j^{(n)}})}{\sum_{B\in\tau_t(n)}\prod_{j=1}^{t}(b_j-1)!\prod_{j=1}^{t}p(y_{B_j})}.$$

Ascolani et al. (2022) prove that

$$C_{\mathrm{DP}}(n,t,s) \to 0 \text{ as } n\to\infty \quad \forall 0<t<s. \tag{4.9}$$

Now, since our expression $R_{\mathrm{PYP}}(n,1,2)$ above does not depend on α, it is identical to the corresponding expression in the setup of Miller and Harrison (2014), who prove that it does not converge to zero as $n\to\infty$. What is left to show is that our expression $C_{\mathrm{PYP}}(n,1,2)$ above does not converge to zero as $n\to\infty$.

We then have,

$$\begin{aligned}
C_{\mathrm{PYP}}(n,1,2) &= \frac{\int \sigma\left(1+\frac{\alpha}{\sigma}\right)\frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}{\int \frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha} \\
&= \sigma + \frac{\int \alpha\frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha}{\int \frac{\zeta(\alpha)}{(1+\alpha)_{(n-1)}}d\alpha} = \sigma + \frac{\int \alpha^2\frac{\zeta(\alpha)}{(\alpha)_{(n)}}d\alpha}{\int \alpha\frac{\zeta(\alpha)}{(\alpha)_{(n)}}d\alpha} \\
&= \sigma + C_{\mathrm{DP}}(n,1,2) \to \sigma \text{ as } n \to \infty,
\end{aligned}$$

where the final line above comes from the special case of Equation (4.9) where $t = 1$ and $s = 2$. □

4.4.3 Proof of Theorem 4.3.0.2

Proof. Throughout this proof, we let E denote expectation with respect to the distribution of the data, $P^\star$ (whose pdf $p^\star$ is given in Equation (4.1)). Recall that we denote the j^{th} cluster of $A^{(n)}$ by $A_j^{(n)}$, and we denote the cardinality of $A_j^{(n)}$ by a_j. We use the notation $[n]$ to represent the cluster of $A^{(n)\star}$, which is of cardinality n. For any cluster, say B of the data, let y_B denote the data in that cluster, and let $p(y_B)$ denote the marginal likelihood of all of the data in that cluster:

$$p(y_B) = \int \prod_{i\in B} k(y_i|\phi)q_0(\phi)d\phi.$$

We define the set $\Omega_n := \{y_{1:n}|Z_n > \frac{1}{2}\}$, where Z_n is defined in Lemma 4.4.1.4. On the set Ω_n, by Lemma 4.4.1.4, it holds that

$$\frac{p^\star(y_{[n]})}{p(y_{[n]})} \leq \frac{n\sqrt{\log(n)}}{WZ_n} \leq \frac{2n\sqrt{\log(n)}}{W}, \tag{4.10}$$

where W is a fixed constant. Therefore, on the set Ω_n, for any partition $A^{(n)}$ with s

clusters, we have

$$\begin{aligned}
P^\star\left(\frac{\Pi_{DP}(A^{(n)}|y_{1:n})}{\Pi_{DP}(A^{(n)\star}|y_{1:n})} > 1\right) &= P^\star\left(\frac{\alpha_n^s}{\alpha_n}\frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{\prod_{j=1}^s p(y_{A_j^{(n)}})}{p(y_{[n]})} > 1\right) \\
&= P^\star\left(\frac{\alpha_n^s}{\alpha_n}\frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{\prod_{j=1}^s p(y_{A_j^{(n)}})}{p^\star(y_{[n]})}\frac{p^\star(y_{[n]})}{p(y_{[n]})} > 1\right) \\
&\leq P^\star\left(\frac{\alpha_n^s}{\alpha_n}\frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{\prod_{j=1}^s p(y_{A_j^{(n)}})}{p^\star(y_{[n]})}\frac{2n\sqrt{\log(n)}}{W} > 1\right) \\
&\leq \alpha_n^{s-1}\frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{2n\sqrt{\log(n)}}{W}E\left(\frac{\prod_{j=1}^s p(y_{A_j^{(n)}})}{p^\star(y_{[n]})}\right) \\
&\leq \alpha_n^{s-1}\frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{2n\sqrt{\log(n)}}{W}\left(\frac{U}{m}\right)^s\prod_{j=1}^s\frac{1}{a_j+1} \\
&\leq \alpha_n^{s-1}\frac{\prod_{j=1}^s (\tilde{a}_j-1)!}{(n-1)!}\frac{2n\sqrt{\log(n)}}{W}\left(\frac{U}{m}\right)^s\prod_{j=1}^s\frac{1}{\tilde{a}_j+1} \\
&= \alpha_n^{s-1}\frac{(s-1)!}{(n-1)!}\frac{2n\sqrt{\log(n)}}{W}\left(\frac{U}{m}\right)^s\prod_{j=1}^s\frac{1}{2^{s-1}(n-s+2)} \\
&= \alpha_n^{s-1}\frac{(s-1)!}{(n-1)!}\frac{2n\sqrt{\log(n)}}{W}\left(\frac{U}{m}\right)^s\frac{1}{2n} \\
&= \frac{(s-1)!}{(n-1)!}\frac{1}{W}\left(\frac{U}{m}\right)^s\alpha_n^{s-1}\sqrt{\log(n)}. \qquad (4.11)
\end{aligned}$$

The first line of the above comes from applying Bayes' rule to the numerator and to the denominator, with the prior distributions on partitions given in Equation (4.2), the third line comes from applying Equation (4.10), the fourth line comes from using the Markov inequality, the fifth line comes from Lemma 4.4.1.5, the sixth line comes from Lemma 4.4.1.2 and Lemma 4.4.1.3, the eighth line comes form the fact that the function $h(s) = \frac{1}{2^{s-1}(n-s+2)}$ is strictly decreasing in s (since $\frac{h(s+1)}{h(s)} = \frac{1}{2}\frac{n-s+2}{n-s+1} < 1\ \forall s < 1$) and thus takes its maximal value at $s = 2$ (recall that we consider values of s in the range $\{2, \ldots, n\}$), and the final line comes from simply rearranging the terms.

Now, by Lemma 4.4.1.6, we have that the MAP partition must respect the ordering of the observations. For $s \in \{1, \ldots, n\}$, let $\mathcal{A}^{(\backslash)}{}_s$ denote the set of partitions that have

s clusters and that respect the ordering of the data. Note that the cardinality of the set $\mathcal{A}^{(\backslash)}{}_s$ is equal to $\frac{(n-1)!}{(n-s)!}$.

We then have,

$$
\begin{aligned}
&P^\star\left(\exists A^{(n)} \neq A^{(n)\star} \text{ such that } \frac{\Pi_{DP}\left(A^{(n)}|y_{1:n}\right)}{\Pi_{DP}\left(A^{(n)\star}|y_{1:n}\right)} > 1\right)\\
&\leq P^\star\left(\left\{\exists A^{(n)} \neq A^{(n)\star} \text{ such that } \frac{\Pi_{DP}\left(A^{(n)}|y_{1:n}\right)}{\Pi_{DP}\left(A^{(n)\star}|y_{1:n}\right)} > 1\right\} \cap \Omega_n\right) + P^\star\left(\Omega_n^c\right)\\
&\leq \sum_{s\in\{2,\ldots,n\}} \sum_{A^{(n)}\in\mathcal{A}^{(\backslash)}{}_s} P^\star\left(\left\{\frac{\Pi_{DP}\left(A^{(n)}|y_{1:n}\right)}{\Pi_{DP}\left(A^{(n)\star}|y_{1:n}\right)} > 1\right\} \cap \Omega_n\right) + P^\star\left(\Omega_n^c\right)\\
&\leq \sum_{s\in\{2,\ldots,n\}} \frac{(n-1)!}{(n-s)!}\frac{(s-1)!}{(n-1)!}\frac{1}{W}\left(\frac{U}{m}\right)^s \alpha_n^{s-1}\sqrt{\log(n)} + P^\star\left(\Omega_n^c\right)\\
&\leq \frac{1}{W}\left(\frac{U}{m}\right)^2 \alpha_n \sum_{i=0}^{\infty}\left(\frac{U\alpha_n}{m}\right)^i \sqrt{\log(n)} + P^\star\left(\Omega_n^c\right)\\
&= \frac{U}{mW}\left(\frac{\frac{U\alpha_n}{m}}{1-\frac{U\alpha_n}{m}}\right)\sqrt{\log(n)} + P^\star\left(\Omega_n^c\right)\\
&\to 0
\end{aligned}
$$

as $n \to \infty$. The third line of the above comes from Lemma 4.4.1.6, the fourth line comes from applying Equation 4.11, and the final line comes from the fact that $\alpha_n = o\left(\frac{1}{\log(n)}\right)$ combined by the fact that By Lemma 4.4.1.4, $P^\star\left(\Omega_n^c\right) \to 0$ as $n \to \infty$.

□

4.5 Discussion

In our Theorem 4.3.0.1 we have proved inconsistency for the number of clusters when fitting single-component mixtures with Pitman–Yor mixture models with a prior on the concentration parameter α and fixed discount parameter σ. Our result holds when the true number of clusters in the data-generating mechanism is one. While hinting at what to expect, further study would be needed to fully understand clustering consistency for a data-generating mechanism with an arbitrary number of components. While our result is limited to the setting where the discount parameter σ is kept fixed, it is common in practice to put a prior on both PY parameters α and σ in PY mixture models. We carried out a short simulation study (see Section 4.6.2), which suggests inconsistency in this case, but consistency when keeping α fixed and putting a prior on σ. Both situations are the subject of current investigation.

In our Theorem 4.3.0.2, we have proved the posterior consistency for the MAP partition when fitting a single-component mixture with Dirichlet process mixture models when concentration parameter α_n goes to zero as the number of obersvations goes to infinity. This result seems to generalise to multi-component mixtures, in particular when one imposes strong assumptions on the separability of the mixture components, as is done by Ascolani et al. (2022). This is a topic of ongoing work.

Our result contrasts the inconsistency result of Rajkowski (2019) who consider the concentration parameter α to be a fixed constant, highlighting the crucial impact of the treatment of α when considering consistency. While it is not uncommon to choose α_n to be a decreasing sequence, it would be arguably more natural in a Bayesian setting to put a prior on α, as is done by Ascolani et al. (2022). It would be interesting to investigate if consistency of the MAP partition holds in such a setting.

The MAP is a very basic estimator and has been criticized by some authors as not being optimal in the context of partitioning data (Wade and Ghahramani (2018)). Indeed, a clustering of the data that differs from the true clustering by just one data point is assigned the same loss as a clustering of the data that is completely different. An interesting future research question could be the asymptotic behavior of other clustering estimators, such as the estimator that minimizes Binder's loss or the variation of information (VI) loss introduced in Wade and Ghahramani (2018).

4.6 Appendix

4.6.1 Proof of Lemmas

Proof of Lemma 4.4.1.2

Proof. Suppose that at least two of the $\tilde{a}_j$ are greater than one. Without loss of generality, $\tilde{a}_1 \neq 1$ and $\tilde{a}_2 \neq 1$ with $\tilde{a}_1 \geq \tilde{a}_2$. Now let $\{\hat{a}_j\}_{j=1}^s$ be another set of strictly positive integers satisfying $\sum_{j=1}^s \hat{a}_j = n$, defined by $\hat{a}_1 = \tilde{a}_1 + 1$, $\hat{a}_2 = \tilde{a}_2 - 1$, and $\hat{a}_j = \tilde{a}_j$ for $j = 3, \ldots, s$. Then,

$$\frac{\prod_{j=1}^s (\hat{a}_j - 1)!}{\prod_{j=1}^s (\tilde{a}_j - 1)!} = \frac{\tilde{a}_1!}{(\tilde{a}_1 - 1)!} \frac{(\tilde{a}_2 - 2)!}{(\tilde{a}_2 - 1)!} = \frac{\tilde{a}_1}{\tilde{a}_2 - 1} > 1.$$

This contradicts $\{\tilde{a}_j\}_{j=1}^s$ being the set of integers maximising $\prod_{j=1}^s (a_j - 1)!$. We conclude that $\tilde{a}_j = 1$ for all except one cluster. □

Proof of Lemma 4.4.1.3

Proof. Suppose that at least two of the $\tilde{a}_j$ are greater than one. Without loss of generality, $\tilde{a}_1 \neq 1$ and $\tilde{a}_2 \neq 1$ with $\tilde{a}_1 \geq \tilde{a}_2$. Now let $\{\hat{a}_j\}_{j=1}^s$ be another set of strictly positive integers satisfying $\sum_{j=1}^s \hat{a}_j = n$, defined by $\hat{a}_1 = \tilde{a}_1 + 1$, $\hat{a}_2 = \tilde{a}_2 - 1$, and $\hat{a}_j = \tilde{a}_j$ for $j = 3, \ldots, s$. Then,

$$\frac{\prod_{j=1}^s (\hat{a}_j + 1)}{\prod_{j=1}^s (\tilde{a}_j + 1)} = \left(\frac{\tilde{a}_1 + 2}{\tilde{a}_1 + 1}\right) \frac{\tilde{a}_2}{\tilde{a}_2 + 1} < 1,$$

since for any $a < b$, one has $\frac{a}{a+1} < \frac{b}{b+1}$ (here, $a = \tilde{a}_2$ and $b = \tilde{a}_1 + 1$). This contradicts $\{\tilde{a}_j\}_{j=1}^s$ being the set of integers minimising $\prod_{j=1}^s (a_j + 1)$. We conclude that $\tilde{a}_j = 1$ for all except one cluster. □

4.6.2 Simulation studies

Single-component mixture

We illustrate Theorem 4.3.0.1 using data generated from the following single-component Gaussian location "mixture" model

$$p^\star(y) = \mathcal{N}(y|\mu, \Sigma)$$

where $\mu = (0.8, 0.8)$ and $\Sigma = 0.05\, I_2$. We adapt the Importance Conditional Sampler for PYP mixtures of Canale et al. (2022), using the same prior specification on μ and Σ as Malsiner-Walli et al. (2016), provided below.

$$\begin{gathered} G \sim \text{PYP}(\alpha, \sigma, Q_0), \quad \mu \sim \mathcal{N}(b_0, B_0) \\ \Sigma^{-1} \sim \mathcal{W}(c_0, C_0), \quad C_0 \sim \mathcal{W}(Q_0, Q_0) \end{gathered}$$

We put a Gamma(200, 20) prior on the concentration parameter α, and keep the discount parameter constant at $\sigma = 0.5$. The Gamma prior on α satisfies the conditions of our proof. We consider data sets of size $n \in \{50, 200, 500, 2000\}$.

Figure 4.1 illustrates the result of our theorem, showing inconsistency of the number of clusters under this set-up.

Multi-component mixture

We investigate a possible extension of Theorem 4.3.0.1 where the number of components may be greater than one using data generated from a Gaussian location mixture with $t = 3$ components,

$$f(y) = \sum_{i=1}^{3} p_i \, \mathcal{N}(y|\mu_i, \Sigma),$$

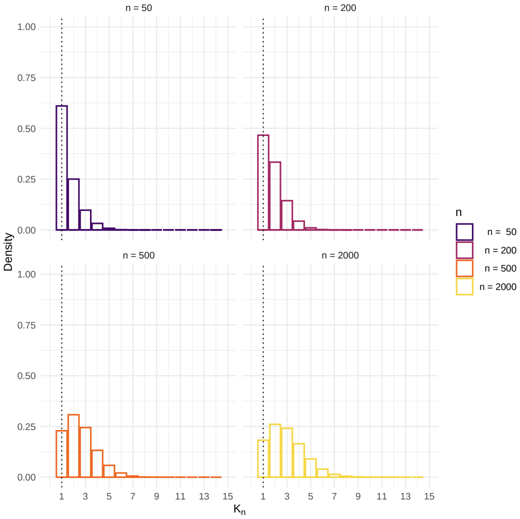
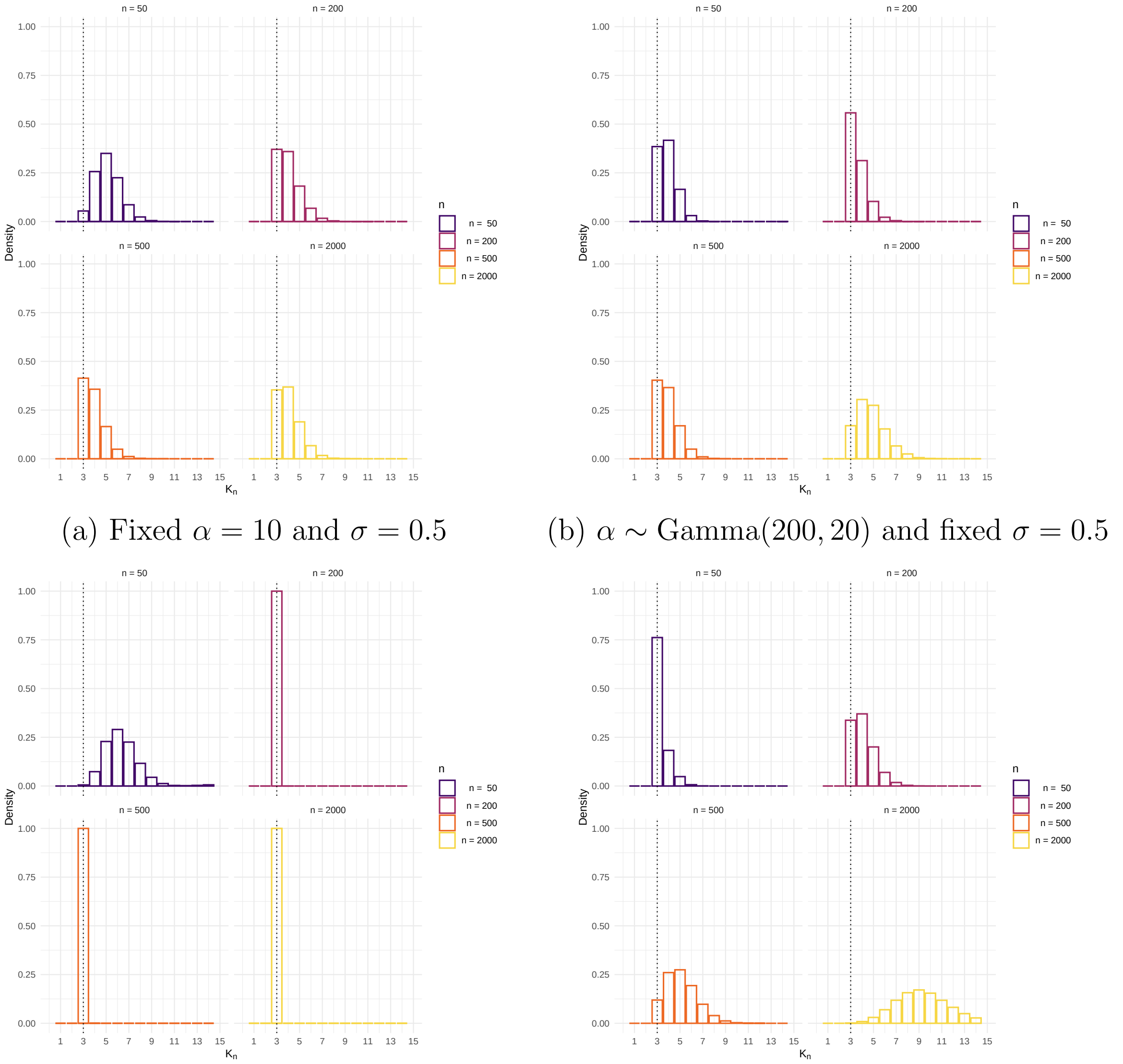


Figure 4.1: Posterior distribution of the number of clusters K_n under a Pitman–Yor process single-component mixture for various choices of n and with $\alpha \sim \text{Gamma}(200, 20)$ and fixed $\sigma = 0.5$ (i.e. under the set-up of our Theorem 4.3.0.1). We observe that the posterior does not concentrate on the true value $K_n = 1$.

where $p = (p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, $\mu_1 = (0.8, 0.8)$, $\mu_2 = (0.8, -0.8)$, $\mu_3 = (-0.8, 0.8)$ and $\Sigma = 0.05I_2$. We use a similar algorithm and prior specification to that used in the single-component case.

Figure 4.2 (a) illustrates the result proved in Miller and Harrison (2014), that Pitman–Yor mixture models with fixed parameters α and σ are inconsistent for the number of clusters. Figure 4.2 (b), (c) and (d) illustrate cases not covered by current theoretical results. Figure 4.2 (b) shows inconsistency in the case where the parameter σ is fixed but when a prior is put on the parameter α. This suggests that our Theorem 4.3.0.1 may generalise to multi-component mixture models. Figure 4.2 (c) shows consistency in the case when the parameter α is fixed and a prior is put on the parameter σ. Figure 4.2 (d) shows inconsistency in the case where a prior is put on both parameters α and σ.



(a) Fixed $\alpha = 10$ and $\sigma = 0.5$

(b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed $\sigma = 0.5$

(c) Fixed $\alpha = 10$ and $\sigma \sim \text{Unif}(0, 1)$

(d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Unif}(0, 1)$

Figure 4.2: Posterior distribution of the number of clusters K_n under a Pitman–Yor process mixture for various choices of n and with (a) fixed parameters α and σ; (b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed σ; (c) fixed α and $\sigma \sim \text{Unif}(0, 1)$; and (d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Unif}(0, 1)$, where the true data-generating process has 3 components.

Bibliography

Alamichel, L., D. Bystrova, J. Arbel, and G. K. K. King (2022). Bayesian mixture models (in) consistency for the number of clusters. *arXiv preprint arXiv:2210.14201*.

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

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). Clustering consistency with Dirichlet process mixtures. *arXiv proprietor arXiv:2205.12924*.

Attorre, F., V. E. Cambria, E. Agrillo, N. Alessi, M. Alfò, M. De Sanctis, L. Malatesta, T. Sitzia, R. Guarino, C. Marcenò, et al. (2020). Finite mixture model-based classification of a complex vegetation system. *Vegetation Classification and Survey 1*, 77–86.

Baker, J. K. (1979). Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America 65*(S1), S132–S132.

Barron, A., M. J. Schervish, and L. Wasserman (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics 27*(2), 536–561.

Beal, M., Z. Ghahramani, and C. Rasmussen (2001). The infinite hidden Markov model. *Advances in Neural Information Processing Systems 14*.

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems 1*, 577–584.

Bell, E. T. (1934). Exponential polynomials. *Annals of Mathematics*, 258–277.

Bernardo, J. M. and A. F. Smith (2009). *Bayesian theory*, Volume 405. John Wiley & Sons.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika 65*(1), 31–38.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*(2), 353–355.

Blei, D. M., M. I. Jordan, et al. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis 1*(1), 121–144.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Blum, M. G. (2010). Approximate Bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association 105*(491), 1178–1187.

Blum, M. G. B. and O. François (2009). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing 20*, 63–73.

Bortot, P., S. G. Coles, and S. A. Sisson (2007). Inference for stereological extremes. *Journal of the American Statistical Association 102*(477), 84–92.

Broët, P., S. Richardson, and F. Radvanyi (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology 9*(4), 671–683.

Canale, A., R. Corradin, and B. Nipoti (2022, May). Importance conditional sampling for Pitman–Yor mixtures. *Statistics and Computing 32*(3), 40.

Charniak, E. (1996). Tree-bank grammars. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1031–1036.

Chatain, K., P. Schlenker, D. Demolin, F. Mendes, R. J. Ryder, and E. Chemla (2023). Muriqui monkey vocalizations are complex: a formal and a statistical assessment. *Under revision*.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory 2*(3), 113–124.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika 89*(3), 539–552.

Chopin, N., O. Papaspiliopoulos, et al. (2020). *An introduction to sequential Monte Carlo*, Volume 4. Springer.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics 29*(4), 589–637.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics 4*, 201–218.

Dahl, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Analysis 4*(2), 243–264.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 212–229.

Demolin, D., C. Ades, and F. Mendes (2016). Context-sensitive grammars in Muriqui vocalizations. *Scientific Reports*.

Doob, J. L. (1949). Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications*, 23–27.

Doucet, A., N. De Freitas, N. J. Gordon, et al. (2001). *Sequential Monte Carlo methods in practice*, Volume 1. Springer.

Duffie, D. and K. J. Singleton (1990). Simulated moments estimation of Markov models of asset prices. Technical report, National Bureau of Economic Research.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Escobar, M. D. and M. West (1998). Computing nonparametric hierarchical models. *Practical Nonparametric and Semiparametric Bayesian Statistics*, 1–22.

Fearnhead, P. and D. Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), 419–474.

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

Finkel, J. R., T. Grenager, and C. D. Manning (2007). The infinite tree. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 272–279.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Franssen, S. and A. van der Vaart (2022). Bernstein-von Mises theorem for the Pitman-Yor process of nonnegative type. *Electronic Journal of Statistics 16*(2), 5779–5811.

Frazier, D. T., G. M. Martin, C. P. Robert, and J. Rousseau (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika 105*(3), 593–607.

Fruhwirth-Schnatter, S., G. Celeux, and C. P. Robert (2019). *Handbook of mixture analysis*. CRC press.

Frühwirth-Schnatter, S., C. Pamminger, A. Weber, and R. Winter-Ebmer (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics 27*(7), 1116–1137.

Fuentes-García, R., R. H. Mena, and S. G. Walker (2019). Modal posterior clustering motivated by Hopfield's network. *Computational Statistics & Data Analysis 137*, 92–100.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al. (2002). The structure of haplotype blocks in the human genome. *Science 296*(5576), 2225–2229.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, 1317–1339.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood.

Ghosal, L. and A. van der Vaart (2017). Fundamentals of Bayesian nonparametric inference.

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

Ghosal, S., J. K. Ghosh, and A. W. Van Der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 500–531.

Ghosal, S. and A. Van Der Vaart (2007). Convergence rates of posterior distributions for non iid observations. *Annals of Statistics 35*(1), 192–223.

Gilks, W. R. and C. Berzuini (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(1), 127–146.

Goldwater, S. and T. Griffiths (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*, pp. 744–751.

Goldwater, S., T. L. Griffiths, and M. Johnson (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 673–680.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*(S1), S85–S118.

Grelaud, A., J.-M. Marin, C. P. Robert, F. Rodolphe, and J.-F. Taly (2009). Abc likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis 4*(2), 317–335.

Grün, B. (2019). Model-based clustering. In *Handbook of mixture analysis*, pp. 157–192. CRC Press, Taylor & Francis Group.

Guha, A., N. Ho, and X. Nguyen (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli 27*(4), 2159–2188.

Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

Johnson, M., T. L. Griffiths, and S. Goldwater (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 139–146.

Joshi, A. K., S. R. Kosaraju, and H. Yamada (1969). String adjunct grammars. In *10th Annual Symposium on Switching and Automata Theory (swat 1969)*, pp. 245–262. IEEE.

Kim, S., M. G. Tadesse, and M. Vannucci (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika 93*(4), 877–893.

Kruijer, W., J. Rousseau, and A. Van Der Vaart (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics 4*, 1225–1257.

Laplace, P.-S. (1810). Mémoire sur les intégrales définies et leur application aux probabilités, et spécialementa la recherche du milieu qu'il faut choisir entre les résultats des observations. *Mem. Acad. Sci.(I), XI, Section V*, 375–387.

Lari, K. and S. J. Young (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech & Language 4*(1), 35–56.

Lartillot, N. and H. Philippe (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution 21*(6), 1095–1109.

Lau, J. W. and P. J. Green (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics 16*(3), 526–558.

Li, J., S. Ray, and B. G. Lindsay (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research 8*(8).

Li, W. and P. Fearnhead (2018a). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika 105*(2), 301–318.

Li, W. and P. Fearnhead (2018b). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika 105*(2), 285–299.

Liang, P., S. Petrov, M. I. Jordan, and D. Klein (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 joint conference on empirical* *methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 688–697.

Lijoi, A., I. Prünster, and S. G. Walker (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association 100*(472), 1292–1296.

Lijoi, A., I. Prünster, and S. G. Walker (2007). Bayesian consistency for stationary models. *Econometric Theory 23*(4), 749–759.

Liu, J. S., R. Chen, and W. H. Wong (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Association 93*(443), 1022–1031.

Liu, J. S. and J. S. Liu (2001). *Monte Carlo strategies in scientific computing*, Volume 75. Springer.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 351–357.

MacEachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*(2), 223–238.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297. University of California Los Angeles LA USA.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing 26*(1-2), 303–324.

Manning, C. and H. Schutze (1999). *Foundations of statistical natural language processing*. MIT Press.

Marin, J.-M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012). Approximate Bayesian computational methods. *Statistics and Computing 22*(6), 1167–1180.

Medvedovic, M. and S. Sivaganesan (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics 18*(9), 1194–1206.

Medvedovic, M., P. Succop, R. Shukla, and K. Dixon (2001). Clustering mutational spectra via classification likelihood and Markov chain Monte Carlo algorithms. *Journal of Agricultural, Biological, and Environmental Statistics 6*, 19–37.

Metropolis, N. and S. Ulam (1949). The Monte Carlo method. *Journal of the American Statistical Association 44*(247), 335–341.

Miller, J. W. and M. T. Harrison (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pp. 199–206.

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

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine 22*(19), 3055–3071.

Müller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika 83*(1), 67–79.

Müller, P., F. A. Quintana, and G. Page (2018). Nonparametric Bayesian inference in applications. *Statistical Methods & Applications 27*, 175–206.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing 11*, 125–139.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics 41*(1), 370–400.

Ohn, I. and L. Lin (2023). Optimal Bayesian estimation of Gaussian mixtures with growing number of components. *Bernoulli 29*(2), 1195–1218.

Onogi, A., M. Nurimoto, and M. Morita (2011). Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics 12*(1), 1–16.

Perman, M., J. Pitman, and M. Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields 92*(1), 21–39.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields 102*(2), 145–158.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.

Pollard, C. J. (1984). Generalized phrase structure grammars, head grammars, and natural language. *PhD dissertation, Stanford University*.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*(12), 1791–1798.

Rajkowski, Ł. (2019). Analysis of the maximal a posteriori partition in the Gaussian Dirichlet process mixture model. *Bayesian Analysis 14*(2), 477–494.

Ramírez, V. M., F. Forbes, J. Arbel, A. Arnaud, and M. Dojat (2019). Quantitative MRI characterization of brain abnormalities in de novo Parkinsonian patients. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1572–1575. IEEE.

Ratmann, O., O. Jørgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of h. pylori and p. falciparum. *PLoS Comput Biol 3*(11), e230.

Ray, K. and A. van der Vaart (2021). On the Bernstein-von Mises theorem for the Dirichlet process. *Electronic Journal of Statistics 15*, 2224–2246.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Volume 2. Springer.

Robert, C. P., G. Casella, and G. Casella (1999). *Monte Carlo statistical methods*, Volume 2. Springer.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rousseau, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and its Application 3*, 211–231.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*(5), 689–710.

Ryder, R. J., L. Murray, J. Rousseau, and A. Thin (2023). A Bayesian nonparametric methodology for inferring grammar complexity. *In preparation*.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 4*(1), 10–26.

Seki, H., T. Matsumura, M. Fujii, and T. Kasami (1991). On multiple context-free grammars. *Theoretical Computer Science 88*(2), 191–229.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 639–650.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In *Philosophy, language, and artificial intelligence*, pp. 79–89. Springer.

Sisson, S. A., Y. Fan, and M. Beaumont (2018). *Handbook of approximate Bayesian computation*. CRC Press.

Tanaka, M. M., A. R. Francis, F. Luciani, and S. Sisson (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics 173*(3), 1511–1520.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from dna sequence data. *Genetics 145*(2), 505–518.

Teh, Y., M. Jordan, M. Beal, and D. Blei (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems 17*.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface 6*(31), 187–202.

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis 13*(2), 559–626.

Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika 90*(2), 482–488.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics 32*(5), 2028–2043.

Walker, S. and N. L. Hjort (2001). On Bayesian consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology 63*(4), 811–821.

West, M. and M. D. Escobar (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature 466*(7310), 1102–1104.