Journal of Machine Learning Research 15 (2014) 3333-3370 Submitted 9/13; Revised 5/14; Published 10/14

Inconsistency of Pitman–Yor Process Mixtures for the Number of Components

Jeffrey W. Miller JEFFREY_MILLER@BROWN.EDU
Matthew T. Harrison MATTHEW_HARRISON@BROWN.EDU
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA

Editor: Yee Whye Teh

Abstract

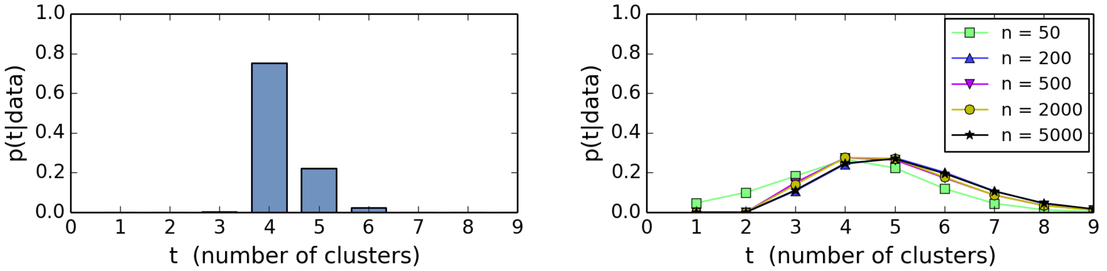
In many applications, a finite mixture is a natural model, but it can be difficult to choose an appropriate number of components. To circumvent this choice, investigators are increasingly turning to Dirichlet process mixtures (DPMs), and Pitman–Yor process mixtures (PYMs), more generally. While these models may be well-suited for Bayesian density estimation, many investigators are using them for inferences about the number of components, by considering the posterior on the number of components represented in the observed data. We show that this posterior is not consistent—that is, on data from a finite mixture, it does not concentrate at the true number of components. This result applies to a large class of nonparametric mixtures, including DPMs and PYMs, over a wide variety of families of component distributions, including essentially all discrete families, as well as continuous exponential families satisfying mild regularity conditions (such as multivariate Gaussians).

Keywords: consistency, Dirichlet process mixture, number of components, finite mixture, Bayesian nonparametrics

1. Introduction

We begin with a motivating example. In population genetics, determining the "population structure" is an important step in the analysis of sampled data. To illustrate, consider the impala, a species of antelope in southern Africa. Impalas are divided into two subspecies: the common impala occupying much of the eastern half of the region, and the black-faced impala inhabiting a small area in the west. While common impalas are abundant, the number of black-faced impalas has been decimated by drought, poaching, and declining resources due to human and livestock expansion. To assist conservation efforts, Lorenzen et al. (2006) collected samples from 216 impalas, and analyzed the genetic variation between/within the two subspecies.

A key part of their analysis consisted of inferring the population structure—that is, partitioning the data into distinct populations, and in particular, determining how many such populations there are. To infer the impala population structure, Lorenzen et al. employed a widely-used tool called STRUCTURE (Pritchard et al., 2000) which, in the simplest version, models the data as a finite mixture, with each component in the mixture corresponding



(a) Posterior for impala data

(b) Posterior for Gaussian mixture data

Figure 1: Estimated DPM posterior distributions of the number of clusters, with concentration parameter 1: (a) For the impala data of Lorenzen et al. ($n = 216$ data points); we use the same base measure as Huelsenbeck and Andolfatto, and our empirical results, shown here, agree with theirs. (b) For data from the three-component univariate Gaussian mixture $\sum_{i=1}^{3} \pi_i \mathcal{N}(x|\mu_i, \sigma_i^2)$ with $\pi = (0.45, 0.3, 0.25)$, $\mu = (4, 6, 8)$, and $\sigma = (1, 0.2, 0.6)$; we use a base measure with the same parameters as Richardson and Green (1997); each plot is the average over 10 independently-drawn data sets. For both (a) and (b), estimates were made via Gibbs sampling (MacEachern, 1994; Neal, 2000), with 10^5 burn-in sweeps and 2×10^5 sample sweeps.

to a distinct population. Structure uses an *ad hoc* method to choose the number of components, but this comes with no guarantees.

Seeking a more principled approach, Pella and Masuda (2006) proposed using a Dirichlet process mixture (DPM). Now, in a DPM, the number of components is infinite with probability 1, and thus the posterior on the number of components is always, trivially, a point mass at infinity. Consequently, Pella and Masuda instead employed the posterior on the number of clusters (that is, the number of components used in generating the data observed so far) for inferences about the number of components. (The terms “component” and “cluster” are often used interchangeably, but we make the following crucial distinction: a component is part of a mixture distribution, while a cluster is the set of indices of data points coming from a given component.) This DPM approach was implemented in a software tool called Structurama (Huelsenbeck and Andolfatto, 2007), and demonstrated on the impala data of Lorenzen et al.; see Figure 1(a).

Structurama has gained acceptance within the population genetics community, and has been used in studies of a variety of organisms, from apples and avocados, to sardines and geckos (Richards et al., 2009; Chen et al., 2009; Gonzalez and Zardoya, 2007; Leaché and Fujita, 2010). Studies such as these can carry significant weight, since they may be used by officials to make informed policy decisions regarding agriculture, conservation, and public health.

More generally, in a number of applications the same scenario has played out: a finite mixture seems to be a natural model, but requires the user to choose the number of components, while a Dirichlet process mixture offers a convenient way to avoid this choice. For nonparametric Bayesian density estimation, DPMs are indeed attractive, since the posterior on the density exhibits nice convergence properties; see Section 1.2. However, in several applications, investigators have drawn inferences from the posterior on the number

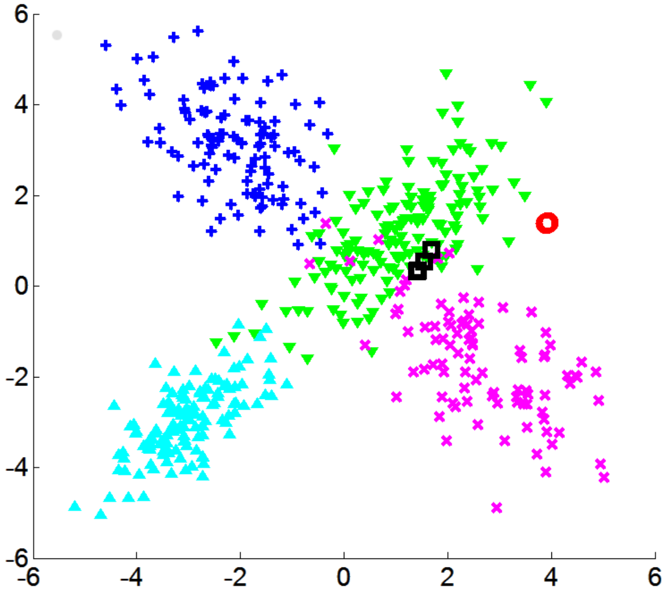


Figure 2: A typical partition sampled from the posterior of a Dirichlet process mixture of bivariate Gaussians, on simulated data from a four-component mixture. Different clusters have different marker shapes (+, ×, ▽, △, ○, □) and different colors. Note the tiny "extra" clusters (○ and □), in addition to the four dominant clusters.

of clusters—not just the density—on the assumption that this is informative about the number of components. Further examples include gene expression profiling (Medvedovic and Sivaganesan, 2002), haplotype inference (Xing et al., 2006), survival analysis (Argiento et al., 2009), econometrics (Otranto and Gallo, 2002), and evaluation of inference algorithms (Fearnhead, 2004). Of course, if the data-generating process is well-modeled by a DPM, then it is sensible to use this posterior for inference about the number of components represented so far in the data—but that does not seem to be the perspective of these investigators, since they measure performance on simulated data coming from finitely many components or populations.

Therefore, it is important to understand the properties of this procedure. Simulation results give some cause for concern; for instance, Figure 1(b) displays results on data from a mixture of univariate Gaussians with three components. The posterior on the number of clusters does not appear to be concentrating as the number of data points n increases. Empirically, it seems that this is because partitions sampled from the posterior often have tiny, transient "extra" clusters (as has been noted before, see Section 1.2); for instance, see Figure 2, showing a typical posterior sample on data from a four-component mixture of bivariate Gaussians. This raises a fundamental question that has not been addressed in the literature: With enough data, will this posterior eventually concentrate at the true number of components? In other words, is it consistent?

It is well-known that under the prior, the number of clusters goes to infinity as $n \to \infty$, with probability 1. However, this does not necessarily imply that the same is true under the posterior—it may be that the likelihood is strong enough to overcome this prior tendency. Of course, in a typical Bayesian setting, the prior is fixed, and as n increases the likelihood overwhelms it. In the present situation, though, both the prior (on the number of clusters) and the likelihood (given the number of clusters) are changing with n, and the resulting behavior of the posterior is far from obvious.

1.1 Overview of Results

In this manuscript, we prove that under fairly general conditions, when using a Dirichlet process mixture, the posterior on the number of clusters will not concentrate at any finite value, and therefore will not be consistent for the number of components in a finite mixture. In fact, our results apply to a large class of nonparametric mixtures including DPMs, and Pitman–Yor process mixtures (PYMs) more generally, over a wide variety of families of component distributions.

Before treating our general results and their prerequisite technicalities, we would like to highlight a few interesting special cases that can be succinctly stated. The terminology and notation used below will be made precise in later sections. To reiterate, our results are considerably more general than the following corollary, which is simply presented for the reader's convenience.

Corollary 1 *Consider a Pitman–Yor process mixture with component distributions from one of the following families:*

(a) $\text{Normal}(\mu, \Sigma)$ *(multivariate Gaussian),*

(b) $\text{Exponential}(\theta)$*,*

(c) $\text{Gamma}(a, b)$*,*

(d) $\text{Log-Normal}(\mu, \sigma^2)$*, or*

(e) $\text{Weibull}(a, b)$ *with fixed shape* $a > 0$*,*

along with a base measure that is a conjugate prior of the form in Section 5.2, or

(f) any discrete family $\{P_\theta\}$ *such that* $\bigcap_\theta \{x : P_\theta(x) > 0\} \neq \varnothing$ *(e.g., Poisson, Geometric, Negative Binomial, Binomial, Multinomial, etc.),*

along with any continuous base measure. Consider any $t \in \{1, 2, \dots\}$*, except for* $t = N$ *in the case of a Pitman–Yor process with parameters* $\sigma < 0$ *and* $\vartheta = N|\sigma|$*. If* $X_1, X_2, \dots$ *are i.i.d. from a mixture with* t *components from the family used in the model, then the posterior on the number of clusters* T_n *is not consistent for* t*, and in fact,*

$$\limsup_{n\to\infty} p(T_n = t \mid X_{1:n}) < 1$$

with probability 1.

This is implied by Theorems 6, 7, and 11. These more general theorems apply to a broad class of partition distributions, handling Pitman–Yor processes as a special case, and they apply to many other families of component distributions: Theorem 11 covers a large class of exponential families, and Theorem 7 covers families satisfying a certain boundedness condition on the densities (including any case in which the model and data distributions have one or more point masses in common, as well as many location–scale families with scale bounded away from zero). Dirichlet processes are subsumed as a further special case, being Pitman–Yor processes with parameters $\sigma = 0$ and $\vartheta > 0$. Also, the assumption of i.i.d. data from a finite mixture is much stronger than what is required by these results.

For PYMs with $\sigma \in [0, 1)$ (including DPMs), our results show that $p(T_n = t \mid X_{1:n})$ does not concentrate at any finite value, however, we have not been able to determine the precise limiting behavior of this posterior; the two most plausible outcomes are that it diverges, or stabilizes at some limiting distribution.

Regarding the exception of $t = N$ when $\sigma < 0$ in Corollary 1: posterior consistency at $t = N$ is possible, however, this could only occur if the chosen parameter N just happens to be equal to the actual number of components, t. On the other hand, consistency at any t can (in principle) be obtained by putting a prior on N; see Section 1.2.1 below. In a similar vein, some investigators place a prior on the concentration parameter ϑ in a DPM, or allow ϑ to depend on n; we conjecture that inconsistency can still occur in these cases, but in this paper, we examine only the case of fixed σ and ϑ.

Truncated stick-breaking processes (Ishwaran and James, 2001) are sometimes used to approximate nonparametric models. In a very limited case—see Section 2.1—our results show that on data from a one-component mixture, such a process truncated at two components will be inconsistent for the number of components. It seems likely that this will extend to truncations at any number of components.

1.2 Discussion / Related Work

We would like to emphasize that this inconsistency should not be viewed as a deficiency of DPMs and PYMs, but is simply due to a misapplication of them. As flexible priors on densities, DPMs are superb, and there are strong results showing that in many cases the posterior on the density converges in L_1 to the true density at the minimax-optimal rate, up to logarithmic factors (Ghosal et al., 1999; Ghosal and Van der Vaart, 2001; Lijoi et al., 2005; Tokdar, 2006; Ghosh and Ghosal, 2006; Tang and Ghosal, 2007; Ghosal and Van der Vaart, 2007; Walker et al., 2007; James, 2008; Wu and Ghosal, 2010; Bhattacharya and Dunson, 2010; Khazaei et al., 2012; Scricciolo, 2012; Pati et al., 2013); for a general overview, see Ghosal (2010).

We would also like to stress that we do not intend to discourage the use of DPMs and PYMs for clustering—provided that the model is indeed well-suited to the application. In some situations, however, it may be that a finite mixture model with an unknown number of components is more appropriate—in particular, for cluster sizes that are all the same order of magnitude—and in such cases, one would expect to get better clustering results by using a variable-dimension mixture model (see Section 1.2.1 below) rather than a DPM or PYM.

Existing work on posterior consistency of nonparametric mixtures has been primarily focused on the density estimation problem (as mentioned above), although recently, Nguyen (2013) has shown that the DPM posterior on the mixing distribution converges in the Wasserstein metric to the true mixing distribution. These existing results do not necessarily imply consistency for the number of components, since any mixture can be approximated arbitrarily well in these metrics by another mixture with a larger number of components (for instance, by making the weights of the extra components infinitesimally small). There seems to be no prior work on consistency of DPMs or PYMs for the number of components in a finite mixture (aside from Miller and Harrison, 2013, in which we discuss the very special case of a DPM on data from a univariate Gaussian "mixture" with one component of known variance).

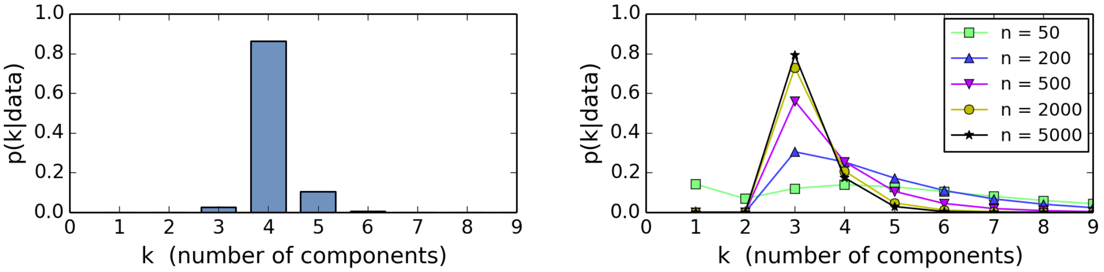
In the context of "species sampling", several authors have studied the Pitman–Yor process posterior (Pitman, 1996; Hansen and Pitman, 2000; James, 2008; Jang et al., 2010; Lijoi et al., 2007, 2008), and interestingly, James (2008) and Jang et al. (2010) have shown that on data from a continuous distribution, the posterior of a Pitman–Yor process with $\sigma > 0$ is inconsistent in the sense that it does not converge weakly to the true distribution. (In contrast, the Dirichlet process is consistent in this sense.) However, this is very different from our situation—in a species sampling model, the observed data is drawn directly from a discrete measure with a Pitman–Yor process prior, while in a PYM model, the observed data is drawn from a mixture with such a measure as the mixing distribution.

Rousseau and Mengersen (2011) proved an interesting result on "overfitted" mixtures, in which data from a finite mixture is modeled by a finite mixture with too many components. In cases where this approximates a DPM, their result implies that the posterior weight of the extra components goes to zero. In a rough sense, this is complementary to our results, which involve showing that there are always some nonempty (but perhaps small) extra clusters.

Empirically, many investigators have noticed that the DPM posterior tends to overestimate the number of components (e.g., West et al., 1994; Ji et al., 2010; Argiento et al., 2009; Lartillot and Philippe, 2004; Onogi et al., 2011, and others), and such observations are consistent with our theoretical results. This overestimation seems to occur because there are typically a few tiny "extra" clusters, and among researchers using DPMs for clustering, this is an annoyance that is sometimes dealt with by pruning such clusters—that is, by removing them before calculating statistics such as the number of clusters (e.g., West et al., 1994; Fox et al., 2007). It may be possible to obtain consistent estimators in this way, but this remains an open question; Rousseau and Mengersen's (2011) results may be applicable here. Other possibilities are using a maximum *a posteriori* (MAP) partition or posterior "mean" partition (Dahl, 2006; Huelsenbeck and Andolfatto, 2007; Onogi et al., 2011) to estimate the number of components; again, the consistency of such approaches remains an open question to our knowledge.

1.2.1 Estimating the Number of Components

A variety of methods for estimating the number of components in a finite mixture have been developed, and many of them come with guarantees of consistency (Henna, 1985; Keribin,



(a) Posterior for impala data

(b) Posterior for Gaussian mixture data

Figure 3: Estimated posterior distributions of the number of components for variable-dimension mixture models applied to the same data sets as in Figure 1. The same priors on component parameters (base measures) were used as in the DPM models.

2000; Nobile, 1994; Leroux, 1992; Ishwaran et al., 2001; James et al., 2001; Henna, 2005; Woo and Sriram, 2006, 2007).

From the Bayesian perspective, perhaps the most natural approach is simply to take a finite mixture model and put a prior on the number of components. For instance, draw the number of components k from a prior which is positive on all positive integers (so there is no *a priori* upper bound), draw mixture weights $(\pi_1, \ldots, \pi_k)$ from, say, a k-dimensional Dirichlet distribution, draw component parameters $\theta_1, \ldots, \theta_k$, and draw the data $X_1, \ldots, X_n$ from the resulting mixture. (Interestingly, it turns out that putting a prior on N in a PYM with $\sigma < 0$ and $\vartheta = N|\sigma|$ is a special case of this; see Gnedin and Pitman, 2006.) Such variable-dimension mixture models have been widely used (Nobile, 1994; Phillips and Smith, 1996; Richardson and Green, 1997; Stephens, 2000; Green and Richardson, 2001; Nobile and Fearnside, 2007), and for density estimation, they have been shown to have posterior rates of concentration similar to Dirichlet process mixtures (Kruijer, 2008; Kruijer et al., 2010). Under the (strong) assumption that the family of component distributions is correctly specified, it has been proven that such models exhibit posterior consistency for the number of components (as well as for the mixing measure and the density) under very general conditions (Nobile, 1994).

Figure 3 shows the posterior on the number of components k for variable-dimension mixture models applied to the same impala data and Gaussian mixture data as in Figure 1. In Figure 3(b), the posterior on k seems to be concentrating at the true number of components (as expected, due to Nobile, 1994), while in Figure 1(b) the DPM posterior does not appear to be concentrating (as expected, due to our results). There is enough information in the data to make the posterior concentrate at the true value; the problem with the DPM posterior is not that estimating the number of components is inherently difficult, but that the DPM posterior is simply the wrong tool for this job.

However, it should be emphasized that this guarantee of posterior consistency for the number of components is contingent upon correct specification of the family of component distributions. In most applications, it seems unreasonable to expect that the data would come from a mixture over a known parametric family, and unfortunately, the posterior on the number of components can be highly sensitive to this type of misspecification—for instance, since any sufficiently regular density can be approximated arbitrarily well by a mixture of Gaussians, if the data distribution is close to but not exactly a finite mixture of Gaussians, a Gaussian mixture model will introduce more and more components as the amount of data increases. It seems that in order to obtain reliable assessments of heterogeneity using mixture models, one needs to carefully consider the effects of potential misspecification. Steps toward addressing this robustness issue have been taken by Woo and Sriram (2006, 2007).

1.3 Idea of the Proof

Roughly speaking, the reason why the posterior on the number of clusters does not concentrate for PYMs with $\sigma \in [0, 1)$ (the $\sigma < 0$ case is somewhat different) is that under the prior, the partition distribution strongly prefers that some of the clusters be very small, and the likelihood is not significantly decreased by splitting off such small clusters. Handling the likelihood—in a general setting—is the challenging part of the proof.

The proof involves showing that $p(T_n = t + 1 \mid X_{1:n})$ is at least the same order of magnitude (asymptotically with respect to n) as $p(T_n = t \mid X_{1:n})$. To get the basic idea of why this occurs, write

$$p(T_n = t \mid X_{1:n}) = \frac{p(X_{1:n}, T_n = t)}{p(X_{1:n})} = \frac{1}{p(X_{1:n})} \sum_{A \in \mathcal{A}_t(n)} p(X_{1:n}|A)p(A), \tag{1}$$

where the sum is over all partitions A of $\{1, \ldots, n\}$ into t parts.

Now, given some t-part partition A, suppose B is a $(t+1)$-part partition obtained from A by splitting off a single element j to be in its own cluster. For Pitman–Yor processes, $p(B)$ is at least the same order of magnitude as $p(A)/n$. In Section 3, this property is encapsulated in Condition 3, which is simple to check for any closed-form partition distribution.

Similarly, it turns out that typically, for a non-negligible fraction of the elements j, the likelihood $p(X_{1:n}|B)$ is at least the same order of magnitude as $p(X_{1:n}|A)$; in Section 3, this is made precise in Condition 4. This is trivial in discrete cases (see Section 4), and often is easy to show in any particular continuous case, but establishing this condition in a general setting requires some work, and it is this that occupies the bulk of the proof (Section 8 and the appendices).

When both of these conditions are satisfied, we show that in the expression for $p(T_n = t \mid X_{1:n})$ in Equation 1, for each term $p(X_{1:n}|A)p(A)$ there are on the order of n terms $p(X_{1:n}|B)p(B)$ in the corresponding expression for $p(T_n = t+1 \mid X_{1:n})$ that collectively are at least the same order of magnitude as $p(X_{1:n}|A)p(A)$.

1.4 Organization of the Paper

In Section 2, we define the family of partition-based mixture models under consideration, which includes Pitman–Yor and Dirichlet process mixtures as special cases. In Section 3, we state a general inconsistency theorem for partition-based mixtures satisfying certain conditions. In Section 4, we apply the theorem to cases satisfying a certain boundedness condition on the densities, including discrete families as a special case. In Section 5, we introduce notation for exponential families and conjugate priors, and in Section 6, we apply the theorem to cases in which the mixture is over an exponential family satisfying some regularity conditions. The rest of the paper contains proofs of the results described in the previous sections: Section 7 contains the proof of the general theorem and its application to discrete or bounded cases, Section 8 contains the proof of the application to exponential families, and the appendices contain a number of supporting results for the exponential family case.

2. Model Distribution

Our analysis involves two probability distributions: one which is defined by the model, and another which gives rise to the data. In this section, we describe the model distribution.

Building upon the Dirichlet process (Ferguson, 1973; Blackwell and MacQueen, 1973; Antoniak, 1974), Dirichlet process mixtures were first studied by Antoniak (1974), Berry and Christensen (1979), Ferguson (1983), and Lo (1984), and were later made practical through the efforts of a number of authors (Escobar, 1988; MacEachern, 1994; Escobar and West, 1995; West, 1992; West et al., 1994; Neal, 1992; Liu, 1994; Bush and MacEachern, 1996; MacEachern and Müller, 1998; MacEachern, 1998; Escobar and West, 1998; MacEachern, 1999; Neal, 2000). Pitman–Yor process mixtures (Ishwaran and James, 2001, 2003) are a generalization of DPMs based on the Pitman–Yor process (Perman et al., 1992; Pitman and Yor, 1997), also known as the two-parameter Poisson–Dirichlet process. We consider a general class of partition-based mixture models that includes DPMs and PYMs.

2.1 Partition Distribution

We will use $p(\cdot)$ to denote probabilities and probability densities under the model. Our model specification begins with a distribution on partitions, or more precisely, on *ordered* partitions. Given $n \in \{1, 2, \dots\}$ and $t \in \{1, \dots, n\}$, let $\mathcal{A}_t(n)$ denote the set of all ordered partitions $A = (A_1, \dots, A_t)$ of $\{1, \dots, n\}$ into t nonempty sets (or "parts"). In other words,

$$\mathcal{A}_t(n) = \Big\{(A_1, \dots, A_t) : A_1, \dots, A_t \text{ are disjoint}, \bigcup_{i=1}^{t} A_i = \{1, \dots, n\},\ |A_i| \geq 1\ \forall i\Big\}.$$

For each $n \in \{1, 2, \dots\}$, consider a probability mass function (p.m.f.) $p(A)$ on $\bigcup_{t=1}^{n} \mathcal{A}_t(n)$. This induces a distribution on t in the natural way, via $p(t \mid A) = I(A \in \mathcal{A}_t(n))$. (Throughout, we use I to denote the indicator function: $I(E)$ is 1 if E is true, and 0 otherwise.) It follows that $p(A) = p(A, t)$ when $A \in \mathcal{A}_t(n)$.

Although it is more common to use a distribution on *unordered* partitions $\{A_1, \dots, A_t\}$, for our purposes it is more convenient to work with the corresponding distribution on ordered partitions $(A_1, \dots, A_t)$ obtained by uniformly permuting the parts. This does not affect the distribution of t. Thus, often, $p(A)$ is invariant under permutations of the parts, but we do not require this. (Also, we do not assume that, as n varies, the sequence of partition distributions necessarily satisfies the marginalization property referred to as "consistency in distribution"; Pitman, 2006.)

For example, the partition distribution for the Dirichlet process is

$$p(A) = \frac{\vartheta^t}{\vartheta_{n\uparrow 1}\; t!} \prod_{i=1}^{t} (|A_i| - 1)! \tag{2}$$

for $A \in \mathcal{A}_t(n)$, where $\vartheta > 0$ and $x_{n\uparrow\delta} = x(x+\delta)(x+2\delta)\cdots(x+(n-1)\delta)$, with $x_{0\uparrow\delta} = 1$ by convention. The $t!$ in the denominator appears since we are working with ordered partitions. More generally, the partition distribution for the Pitman–Yor process is

$$p(A) = \frac{(\vartheta+\sigma)_{t-1\uparrow\sigma}}{(\vartheta+1)_{n-1\uparrow 1}\; t!} \prod_{i=1}^{t} (1-\sigma)_{|A_i|-1\uparrow 1} \tag{3}$$

for $A \in \mathcal{A}_t(n)$, where either $\sigma \in [0,1)$ and $\vartheta \in (-\sigma, \infty)$, or $\sigma \in (-\infty, 0)$ and $\vartheta = N|\sigma|$ for some $N \in \{1, 2, \dots\}$. When $\sigma = 0$, this reduces to the partition distribution of the Dirichlet process. When $\sigma < 0$ and $\vartheta = N|\sigma|$, it is the partition distribution obtained by drawing $q = (q_1, \dots, q_N)$ from a symmetric N-dimensional Dirichlet with parameters $|\sigma|, \dots, |\sigma|$, sampling assignments $Z_1, \dots, Z_n$ i.i.d. from q, and removing any empty parts (Gnedin and Pitman, 2006). Thus, in this latter case, t is always in $\{1, \dots, N\}$.

Stick-breaking processes truncated at N components are sometimes used to approximate nonparametric models (Ishwaran and James, 2001). This approach gives rise to a partition distribution as follows: let $V_i \sim \mathrm{Beta}(a_i, b_i)$ independently for $i = 1, \dots, N-1$, and $V_N = 1$, set $q_i = V_i \prod_{j<i}(1 - V_j)$ for $i = 1, \dots, N$, sample assignments $Z_1, \dots, Z_n$ i.i.d. from q, and remove any empty parts. In general, it seems that this partition distribution takes a slightly complicated form, however, in the very special case when $N = 2$ and $a_1 = b_1$, it is simply a Pitman–Yor process with $\sigma = -a_1 = -b_1$ and $\vartheta = 2|\sigma|$.

2.2 Partition-based Mixture Model

Consider the hierarchical model

$$\begin{aligned} p(A, t) &= p(A) \\ p(\theta_{1:t} \mid A, t) &= \prod_{i=1}^{t} \pi(\theta_i) \\ p(x_{1:n} \mid \theta_{1:t}, A, t) &= \prod_{i=1}^{t} \prod_{j \in A_i} p_{\theta_i}(x_j) \end{aligned} \tag{4}$$

where π is a prior density on component parameters $\theta \in \Theta \subset \mathbb{R}^k$ for some k, and $\{p_\theta : \theta \in \Theta\}$ is a parameterized family of densities on $x \in \mathcal{X} \subset \mathbb{R}^d$ for some d. Here, $x_{1:n} = (x_1, \dots, x_n)$ with $x_i \in \mathcal{X}$, $\theta_{1:t} = (\theta_1, \dots, \theta_t)$ with $\theta_i \in \Theta$, and $A \in \mathcal{A}_t(n)$. Assume that π is a density with respect to Lebesgue measure, and that $\{p_\theta : \theta \in \Theta\}$ are densities with respect to some sigma-finite Borel measure λ on $\mathcal{X}$, such that $(\theta, x) \mapsto p_\theta(x)$ is measurable. (The distribution of x under $p_\theta(x)$ may be discrete, continuous, or neither, depending on λ.)

For $x_1, \dots, x_n \in \mathcal{X}$ and $J \subset \{1, \dots, n\}$, define the *single-cluster marginal*,

$$m(x_J) = \int_\Theta \Big(\prod_{j \in J} p_\theta(x_j) \Big) \pi(\theta)\, d\theta, \tag{5}$$

where $x_J = (x_j : j \in J)$, and assume $m(x_J) < \infty$. By convention, $m(x_J) = 1$ when $J = \varnothing$. Note that $m(x_J)$ is a density with respect to the product measure λ^ℓ on $\mathcal{X}^\ell$, where $\ell = |J|$, and that $m(x_J)$ can (and often will) be positive outside the support of λ^ℓ.

Definition 2 *We refer to such a hierarchical model as a* partition-based mixture model*.*

In particular, it is a *Dirichlet process mixture model* when $p(A)$ is as in Equation 2, or more generally, a *Pitman–Yor process mixture model* when $p(A)$ is as in Equation 3.

The prior on the number of clusters under such a model is $p(T_n = t) = \sum_{A \in \mathcal{A}_t(n)} p(A)$. We use T_n, rather than T, to denote the random variable representing the number of clusters, as a reminder that its distribution depends on n.

Since we are concerned with the posterior $p(T_n = t \mid x_{1:n})$ on the number of clusters, we will be especially interested in the marginal density of $(x_{1:n}, t)$, given by $p(x_{1:n}, T_n = t) = \sum_{A \in \mathcal{A}_t(n)} p(x_{1:n}, A)$. Observe that

$$p(x_{1:n}, A) = p(A) \prod_{i=1}^{t} \int \Big(\prod_{j \in A_i} p_{\theta_i}(x_j) \Big) \pi(\theta_i)\, d\theta_i = p(A) \prod_{i=1}^{t} m(x_{A_i}). \tag{6}$$

For definiteness, we employ the usual version of the posterior of T_n,

$$p(T_n = t \mid x_{1:n}) = \frac{p(x_{1:n}, T_n = t)}{p(x_{1:n})} = \frac{p(x_{1:n}, T_n = t)}{\sum_{t'=1}^{\infty} p(x_{1:n}, T_n = t')}$$

whenever the denominator is nonzero, and $p(T_n = t \mid x_{1:n}) = 0$ otherwise (for notational convenience).

3. General Theorem

The essential ingredients in the main theorem are Conditions 3 and 4 below. For $A \in \mathcal{A}_t(n)$, define $R_A = \bigcup_{i:|A_i| \geq 2} A_i$, and for $j \in R_A$, define $B(A, j)$ to be the element B of $\mathcal{A}_{t+1}(n)$ such that $B_i = A_i \smallsetminus j$ for $i = 1, \ldots, t$, and $B_{t+1} = \{j\}$ (that is, remove j from whatever part it belongs to, and make $\{j\}$ the $(t+1)^{\text{th}}$ part). Let $\mathcal{Z}_A = \{B(A, j) : j \in R_A\}$. For $n > t \geq 1$, define

$$c_n(t) = \frac{1}{n} \max_{A \in \mathcal{A}_t(n)} \max_{B \in \mathcal{Z}_A} \frac{p(A)}{p(B)}, \tag{7}$$

with the convention that $0/0 = 0$ and $y/0 = \infty$ for $y > 0$.

Condition 3 *Assume* $\limsup_{n\to\infty} c_n(t) < \infty$*, given some particular* $t \in \{1, 2, \ldots\}$*.*

For Pitman–Yor processes, Condition 3 holds for all relevant values of t; see Proposition 5 below. Now, given $n \geq t \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$, and $c \in [0, \infty)$, define

$$\varphi_t(x_{1:n}, c) = \min_{A \in \mathcal{A}_t(n)} \frac{1}{n} |S_A(x_{1:n}, c)|$$

where $S_A(x_{1:n}, c)$ is the set of indices $j \in \{1, \ldots, n\}$ such that the part A_ℓ containing j satisfies $m(x_{A_\ell}) \leq c\, m(x_{A_\ell \smallsetminus j}) m(x_j)$.

Condition 4 *Given a sequence of random variables $X_1, X_2, \ldots \in \mathcal{X}$, and $t \in \{1, 2, \ldots\}$, assume*

$$\sup_{c\in[0,\infty)} \liminf_{n\to\infty} \varphi_t(X_{1:n}, c) > 0 \ \text{ with probability } 1.$$

Note that Condition 3 involves only the partition distributions, while Condition 4 involves only the data distribution and the single-cluster marginals.

Proposition 5 *Consider a Pitman–Yor process. If $\sigma \in [0, 1)$ and $\vartheta \in (-\sigma, \infty)$ then Condition 3 holds for any $t \in \{1, 2, \ldots\}$. If $\sigma \in (-\infty, 0)$ and $\vartheta = N|\sigma|$, then it holds for any $t \in \{1, 2, \ldots\}$ except N.*

Proof See Section 7. ■

Theorem 6 *Let $X_1, X_2, \ldots \in \mathcal{X}$ be a sequence of random variables (not necessarily i.i.d.). Consider a partition-based mixture model. For any $t \in \{1, 2, \ldots\}$, if Conditions 3 and 4 hold, then*

$$\limsup_{n\to\infty} p(T_n = t \mid X_{1:n}) < 1 \ \text{ with probability } 1.$$

If, further, the sequence $X_1, X_2, \ldots$ is i.i.d. from a mixture with t components, then with probability 1, the posterior of T_n (under the model) is not consistent for t.

Proof See Section 7. ■

4. Application to Discrete or Bounded Cases

By Theorem 6, the following result implies inconsistency in a large class of PYM models, including essentially all discrete cases (or more generally anything with at least one point mass) and a number of continuous cases as well.

Theorem 7 *Let $X_1, X_2, \ldots \in \mathcal{X}$ be a sequence of random variables (not necessarily i.i.d.). If there exists $U \subset \mathcal{X}$ such that*

(1) $\displaystyle\liminf_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} I(X_j \in U) > 0$ *with probability 1, and*

(2) $\displaystyle\sup\left\{\frac{p_\theta(x)}{m(x)} : x \in U,\, \theta \in \Theta\right\} < \infty$ *(where $0/0 = 0$, $y/0 = \infty$ for $y > 0$),*

then Condition 4 holds for all $t \in \{1, 2, \ldots\}$.

Proof See Section 7. ■

The preceding theorem covers a fairly wide range of cases; here are some examples.

(i) **Finite sample space.** Suppose $\mathcal{X}$ is a finite set, λ is counting measure, and $m(x) > 0$ for all $x \in \mathcal{X}$. Then choosing $U = \mathcal{X}$, Conditions 7(1) and 7(2) of Theorem 7 are trivially satisfied, regardless of the distribution of $X_1, X_2, \ldots$. (Note that when λ is counting measure, $p_\theta(x)$ and $m(x)$ are p.m.f.s on $\mathcal{X}$.) It is often easy to check that $m(x) > 0$ by using the fact that this is true whenever $\{\theta \in \Theta : p_\theta(x) > 0\}$ has nonzero probability under π. This case covers, for instance, Multinomials (including Binomials), and the population genetics model from Section 1.

We should mention a subtle point here: when $\mathcal{X}$ is finite, mixture identifiability might only hold up to a certain maximum number of components (e.g., Teicher, 1963, Proposition 4, showed this for Binomials), making consistency impossible in general—however, consistency might still be possible within that identifiable range. Regardless, our result shows that PYMs are not consistent anyway.

Now, suppose P is a probability measure on $\mathcal{X}$, and $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. Let us abuse notation and write $P(x) = P(\{x\})$ and $\lambda(x) = \lambda(\{x\})$ for $x \in \mathcal{X}$.

(ii) **One or more point masses in common.** If there exists $x_0 \in \mathcal{X}$ such that $P(x_0) > 0$, $\lambda(x_0) > 0$, and $m(x_0) > 0$, then it is easy to verify that Conditions 7(1) and 7(2) are satisfied with $U = \{x_0\}$. (Note that $\lambda(x_0) > 0$ implies $p_\theta(x_0) \le 1/\lambda(x_0)$ for any $\theta \in \Theta$.)

(iii) **Discrete families.** Case (ii) essentially covers all discrete families—e.g., Poisson, Geometric, Negative Binomial, or any power-series distribution (see Sapatinas, 1995, for mixture identifiability of these)—provided that the data is i.i.d.. For, suppose $\mathcal{X}$ is a countable set and λ is counting measure. By case (ii), the theorem applies if there is any $x_0 \in \mathcal{X}$ such that $m(x_0) > 0$ and $P(x_0) > 0$. If this is not so, the model is extremely misspecified, since then the model distribution and the data distribution are mutually singular.

(iv) **Continuous densities bounded on some non-null compact set.** Suppose there exists $c \in (0, \infty)$ and $U \subset \mathcal{X}$ compact such that

   (a) $P(U) > 0$,

   (b) $x \mapsto p_\theta(x)$ is continuous on U for all $\theta \in \Theta$, and

   (c) $p_\theta(x) \in (0, c]$ for all $x \in U$, $\theta \in \Theta$.

Then Condition 7(1) is satisfied due to item (a), and Condition 7(2) follows easily from (b) and (c) since $m(x)$ is continuous (by the dominated convergence theorem) and positive on the compact set U, so $\inf_{x \in U} m(x) > 0$. This case covers, for example, the following families (with any P):

   (a) Exponential(θ), $\mathcal{X} = (0, \infty)$,

   (b) Gamma(a, b), $\mathcal{X} = (0, \infty)$, with variance a/b^2 bounded away from zero,

   (c) Normal(μ, Σ), $\mathcal{X} = \mathbb{R}^d$, (multivariate Gaussian) with $\det(\Sigma)$ bounded away from zero, and

(d) many location–scale families with scale bounded away from zero (for instance, Laplace(μ, σ) or Cauchy(μ, σ), with $\sigma \geq \varepsilon > 0$).

The examples listed in item (iv) are indicative of a deficiency in Theorem 7: Condition 7(2) is not satisfied in some important cases, such as multivariate Gaussians with unrestricted covariance. Nonetheless, it turns out that Condition 4 still holds for many exponential families; we turn to this next.

5. Exponential Families and Conjugate Priors

In this section, we state the usual definitions for exponential families and list the regularity conditions to be assumed.

5.1 Exponential Families

Consider an exponential family of the following form. Fix a sigma-finite Borel measure λ on $\mathcal{X} \subset \mathbb{R}^d$ such that $\lambda(\mathcal{X}) \neq 0$, let $s : \mathcal{X} \to \mathbb{R}^k$ be Borel measurable, and for $\theta \in \Theta \subset \mathbb{R}^k$, define a density p_θ with respect to λ by setting

$$p_\theta(x) = \exp(\theta^{\mathsf{T}} s(x) - \kappa(\theta))$$

where

$$\kappa(\theta) = \log \int_{\mathcal{X}} \exp(\theta^{\mathsf{T}} s(x))\, d\lambda(x).$$

Let P_θ be the probability measure on $\mathcal{X}$ corresponding to p_θ, that is, $P_\theta(E) = \int_E p_\theta(x)\, d\lambda(x)$ for $E \subset \mathcal{X}$ measurable. Any exponential family on $\mathbb{R}^d$ can be written in the form above by reparameterizing if necessary, and choosing λ appropriately. We will assume the following (very mild) regularity conditions.

Condition 8 *Assume the family $\{P_\theta : \theta \in \Theta\}$ is:*

(1) *full, that is,* $\Theta = \{\theta \in \mathbb{R}^k : \kappa(\theta) < \infty\}$*,*

(2) *nonempty, that is,* $\Theta \neq \varnothing$*,*

(3) *regular, that is,* Θ *is an open subset of* $\mathbb{R}^k$*, and*

(4) *identifiable, that is, if* $\theta \neq \theta'$ *then* $P_\theta \neq P_{\theta'}$*.*

Most commonly-used exponential families satisfy Condition 8, including multivariate Gaussian, Exponential, Gamma, Poisson, Geometric, and others. (A notable exception is the Inverse Gaussian, for which Θ is not open.) Let $\mathcal{M}$ denote the *moment space*, that is,

$$\mathcal{M} = \{\mathbb{E}_\theta s(X) : \theta \in \Theta\}$$

where $\mathbb{E}_\theta$ denotes expectation under P_θ. Finiteness of these expectations is guaranteed, thus $\mathcal{M} \subset \mathbb{R}^k$; see Appendix A for this and other well-known properties that we will use.

5.2 Conjugate Priors

Given an exponential family $\{P_\theta\}$ as above, let

$$\Xi = \Big\{(\xi, \nu) : \xi \in \mathbb{R}^k,\ \nu > 0 \text{ s.t. } \xi/\nu \in \mathcal{M}\Big\},$$

and consider the family $\{\pi_{\xi,\nu} : (\xi, \nu) \in \Xi\}$ where

$$\pi_{\xi,\nu}(\theta) = \exp\big(\xi^{\mathrm{T}}\theta - \nu\kappa(\theta) - \psi(\xi,\nu)\big)\, I(\theta \in \Theta)$$

is a density with respect to Lebesgue measure on $\mathbb{R}^k$. Here,

$$\psi(\xi, \nu) = \log \int_\Theta \exp\big(\xi^{\mathrm{T}}\theta - \nu\kappa(\theta)\big)\, d\theta.$$

In Appendix A, we note a few basic properties of this family—in particular, it is a conjugate prior for $\{P_\theta\}$.

Definition 9 *We will say that an exponential family with conjugate prior is* well-behaved *if it takes the form above, satisfies Condition 8, and has $(\xi, \nu) \in \Xi$.*

6. Application to Exponential Families

In this section, we apply Theorem 6 to prove that in many cases, a PYM model using a well-behaved exponential family with conjugate prior will exhibit inconsistency for the number of components.

Condition 10 *Consider an exponential family with sufficient statistics function $s : \mathcal{X} \to \mathbb{R}^k$ and moment space $\mathcal{M}$. Given a probability measure P on $\mathcal{X}$, let $X \sim P$ and assume:*

(1) $\mathbb{E}|s(X)| < \infty$,

(2) $\mathbb{P}(s(X) \in \overline{\mathcal{M}}) = 1$, *and*

(3) $\mathbb{P}(s(X) \in L) = 0$ *for any hyperplane L that does not intersect $\mathcal{M}$.*

Throughout, we use $|\cdot|$ to denote the Euclidean norm. Here, a *hyperplane* refers to a set $L = \{x \in \mathbb{R}^k : x^{\mathrm{T}}y = b\}$ for some $y \in \mathbb{R}^k \smallsetminus \{0\}$, $b \in \mathbb{R}$. In Theorem 11 below, it is assumed that the data comes from a distribution P satisfying Condition 10. In Proposition 12, we give some simple conditions under which, if P is a finite mixture from the exponential family under consideration, then Condition 10 holds.

Theorem 11 *Consider a well-behaved exponential family with conjugate prior (as in Definition 9), along with the resulting collection of single-cluster marginals $m(\cdot)$. Let P be a probability measure on $\mathcal{X}$ satisfying Condition 10 (for the s and $\mathcal{M}$ from the exponential family under consideration), and let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. Then Condition 4 holds for any $t \in \{1, 2, \ldots\}$.*

Proof See Section 7. ■

This theorem implies inconsistency in several important cases. In particular, it can be verified that each of the following is well-behaved (when put in canonical form and given the conjugate prior in Section 5.2) and, using Proposition 12 below, that if P is a finite mixture from the same family then P satisfies Condition 10:

(a) Normal(μ, Σ) (multivariate Gaussian),

(b) Exponential(θ),

(c) Gamma(a, b),

(d) Log-Normal(μ, σ^2), and

(e) Weibull(a, b) with fixed shape $a > 0$.

Combined with the cases covered by Theorem 7, these results are fairly comprehensive.

Proposition 12 *Consider an exponential family $\{P_\theta : \theta \in \Theta\}$ satisfying Condition 8. If $X \sim P = \sum_{i=1}^t \pi_i P_{\theta(i)}$ for some $\theta(1), \ldots, \theta(t) \in \Theta$ and some $\pi_1, \ldots, \pi_t \geq 0$ such that $\sum_{i=1}^t \pi_i = 1$, then*

(1) $\mathbb{E}|s(X)| < \infty$, *and*

(2) $\mathbb{P}(s(X) \in \overline{\mathcal{M}}) = 1$.

If, further, the underlying measure λ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^d$ is open and connected, and the sufficient statistics function $s : \mathcal{X} \to \mathbb{R}^k$ is real analytic (that is, each coordinate function $s_1, \ldots, s_k$ is real analytic), then

(3) $\mathbb{P}(s(X) \in L) = 0$ *for* any *hyperplane $L \subset \mathbb{R}^k$.*

Proof See Appendix A. ■

Sometimes, Condition 10(3) will be satisfied even when Proposition 12 is not applicable. In any particular case, it may be a simple matter to check this condition by using the characterization of $\mathcal{M}$ as the interior of the closed convex hull of support(λs^{-1}) (see Proposition 19(8) in the Appendix).

7. Proof of the General Theorem

The remainder of the paper consists of proofs of the results described in the preceding sections. In this section, we prove Theorem 6, as well as Proposition 5 and the application to discrete or bounded cases in Theorem 7; these proofs do not depend on anything in Section 8 or the appendices.

Proof of Proposition 5 There are two cases: (I) $\sigma \in [0, 1)$ and $\vartheta > -\sigma$, or (II) $\sigma < 0$ and $\vartheta = N|\sigma|$. In either case, we have $1 - \sigma > 0$ and $\vartheta + 1 > 0$; further, $\vartheta + t\sigma > 0$ for (case I) $t \in \{1, 2, \dots\}$, (case II) $t \in \{1, \dots, N-1\}$. Let (case I) $t \in \{1, 2, \dots\}$, (case II) $t \in \{1, \dots, N-1\}$. Let $n > t$, $A \in \mathcal{A}_t(n)$, and $B \in \mathcal{Z}_A$, and suppose $B = B(A, j)$, $j \in A_\ell$. Note that $|A_\ell| \geq 2$.

By the preceding observations, all the factors in the expressions for $p(A)$ and $p(B)$ (Equation 3) are strictly positive, hence

$$\frac{1}{n}\frac{p(A)}{p(B)} = \frac{1}{n}\frac{t+1}{\vartheta + t\sigma}\,(1 - \sigma + |A_\ell| - 2) \leq \frac{t+1}{\vartheta + t\sigma}\,\frac{1 - \sigma + n - 2}{n},$$

which is bounded above for $n \in \{1, 2, \dots\}$. If $t > N$ in case II, then $p(A)/p(B) = 0/0 = 0$ by convention. (If $t = N$ in case II, then $p(A)/p(B) = \infty$.) Therefore, $\limsup_n c_n(t) < \infty$ in either case, for any $t \in \{1, 2, \dots\}$ except $t = N$ in case II. ■

Proof of Theorem 6 The central part of the argument is Lemma 13 below, from which the result follows easily. Let $t \in \{1, 2, \dots\}$, and assume Conditions 3 and 4 hold. Let $x_1, x_2, \dots \in \mathcal{X}$, and suppose $\sup_{c \in [0,\infty)} \liminf_n \varphi_t(x_{1:n}, c) > 0$ (which occurs with probability 1). We show that this implies $\limsup_n p(T_n = t \mid x_{1:n}) < 1$, proving the theorem.

Let $\alpha \in (0, \infty)$ such that $\limsup_n c_n(t) < \alpha$. Choose $c \in [0, \infty)$ and $\varepsilon \in (0, 1)$ such that $\liminf_n \varphi_t(x_{1:n}, c) > \varepsilon$. Choose $N > 2t/\varepsilon$ large enough that for any $n > N$ we have $c_n(t) < \alpha$ and $\varphi_t(x_{1:n}, c) > \varepsilon$. Then by Lemma 13, for any $n > N$,

$$p(T_n = t \mid x_{1:n}) \leq \frac{C_t(x_{1:n}, c)}{1 + C_t(x_{1:n}, c)} \leq \frac{2tc\alpha/\varepsilon}{1 + 2tc\alpha/\varepsilon},$$

since $\varphi_t(x_{1:n}, c) - t/n > \varepsilon - \varepsilon/2 = \varepsilon/2$ (and $y \mapsto y/(1+y)$ is monotone increasing on $[0, \infty)$). Since this upper bound does not depend on n (and is less than 1), $\limsup_n p(T_n = t \mid x_{1:n}) < 1$. ■

Lemma 13 *Consider a partition-based mixture model. Let $n > t \geq 1$, $x_1, \dots, x_n \in \mathcal{X}$, and $c \in [0, \infty)$. If $c_n(t) < \infty$ and $\varphi_t(x_{1:n}, c) > t/n$, then*

$$p(T_n = t \mid x_{1:n}) \leq \frac{C_t(x_{1:n}, c)}{1 + C_t(x_{1:n}, c)},$$

where $C_t(x_{1:n}, c) = t\, c\, c_n(t)/(\varphi_t(x_{1:n}, c) - t/n)$.

Proof To simplify notation, let us denote $\varphi = \varphi_t(x_{1:n}, c)$, $C = C_t(x_{1:n}, c)$, and $S_A = S_A(x_{1:n}, c)$. Recall the definitions of R_A and $B(A, j)$ from the beginning of Section 3. For $A \in \mathcal{A}_t(n)$, note that

$$|R_A \cap S_A| \geq |S_A| - t \geq n\varphi - t > 0. \tag{8}$$

Further, for any $j \in R_A \cap S_A$, we have $p(A) \leq n\, c_n(t)\, p(B(A, j))$ (by the definition of $c_n(t)$, in Equation 7), and $m(x_{A_\ell}) \leq c\, m(x_{A_\ell \smallsetminus j}) m(x_j)$ where A_ℓ is the part containing j (by the definition of $S_A = S_A(x_{1:n}, c)$, in Section 3). Thus, for any $A \in \mathcal{A}_t(n)$, $j \in R_A \cap S_A$, we have (by Equation 6)

$$\begin{aligned} p(x_{1:n}, A) &= p(A) \prod_{i=1}^{t} m(x_{A_i}) \\ &\le n\, c_n(t)\, p(B(A,j))\, c \prod_{i=1}^{t+1} m(x_{B_i(A,j)}) = c\, n\, c_n(t)\, p(x_{1:n}, B(A,j)), \end{aligned}$$

and hence, combining this with Equation 8,

$$\begin{aligned} p(x_{1:n}, A) &\le \frac{c\, n\, c_n(t)}{|R_A \cap S_A|} \sum_{j \in R_A \cap S_A} p(x_{1:n}, B(A,j)) \\ &\le \frac{c\, c_n(t)}{\varphi - t/n} \sum_{B \in \mathcal{A}_{t+1}(n)} p(x_{1:n}, B)\, I(B \in \mathcal{Y}_A), \end{aligned} \tag{9}$$

where $\mathcal{Y}_A = \big\{B(A,j) : j \in R_A \cap S_A\big\}$. For any $B \in \mathcal{A}_{t+1}(n)$,

$$\#\big\{A \in \mathcal{A}_t(n) : B \in \mathcal{Y}_A\big\} \le t, \tag{10}$$

since there are only t parts that B_{t+1} could have come from. Therefore,

$$\begin{aligned} p(x_{1:n}, T_n = t) &= \sum_{A \in \mathcal{A}_t(n)} p(x_{1:n}, A) \\ &\overset{\text{(a)}}{\le} \frac{c\, c_n(t)}{\varphi - t/n} \sum_{A \in \mathcal{A}_t(n)} \sum_{B \in \mathcal{A}_{t+1}(n)} p(x_{1:n}, B)\, I(B \in \mathcal{Y}_A) \\ &= \frac{c\, c_n(t)}{\varphi - t/n} \sum_{B \in \mathcal{A}_{t+1}(n)} p(x_{1:n}, B)\, \#\big\{A \in \mathcal{A}_t(n) : B \in \mathcal{Y}_A\big\} \\ &\overset{\text{(b)}}{\le} \frac{t\, c\, c_n(t)}{\varphi - t/n} \sum_{B \in \mathcal{A}_{t+1}(n)} p(x_{1:n}, B) \\ &= C\, p(x_{1:n}, T_n = t+1), \end{aligned}$$

where (a) is by Equation 9, and (b) is by Equation 10.

If $p(T_n = t \mid x_{1:n}) = 0$, then trivially $p(T_n = t \mid x_{1:n}) \le C/(C+1)$. On the other hand, if $p(T_n = t \mid x_{1:n}) > 0$, then $p(x_{1:n}, T_n = t) > 0$, and therefore

$$\begin{aligned} p(T_n = t \mid x_{1:n}) &= \frac{p(x_{1:n}, T_n = t)}{\sum_{t'=1}^{\infty} p(x_{1:n}, T_n = t')} \\ &\le \frac{p(x_{1:n}, T_n = t)}{p(x_{1:n}, T_n = t) + p(x_{1:n}, T_n = t+1)} \le \frac{C}{C+1}. \end{aligned}$$

■

Proof of Theorem 7 Suppose $U \subset \mathcal{X}$ satisfies (1) and (2), and let $t \in \{1, 2, \dots\}$. Define $c = \sup\big\{\frac{p_\theta(x)}{m(x)} : x \in U,\, \theta \in \Theta\big\}$. Let $n > t$ and $x_1, \dots, x_n \in \mathcal{X}$. Now, for any $x \in U$ and $\theta \in \Theta$, we have $p_\theta(x) \le c\, m(x)$. Hence, for any $J \subset \{1, \dots, n\}$, if $j \in J$ and $x_j \in U$ then

$$m(x_J) = \int_\Theta p_\theta(x_j)\Big[\prod_{i \in J \smallsetminus j} p_\theta(x_i)\Big]\pi(\theta)\, d\theta \le c\, m(x_j) m(x_{J \smallsetminus j}). \tag{11}$$

Thus, letting $R(x_{1:n}) = \big\{j \in \{1, \dots, n\} : x_j \in U\big\}$, we have $R(x_{1:n}) \subset S_A(x_{1:n}, c)$ for any $A \in \mathcal{A}_t(n)$, and hence, $\varphi_t(x_{1:n}, c) \ge \frac{1}{n}|R(x_{1:n})|$.

Therefore, by (1), with probability 1,

$$\liminf_{n\to\infty} \varphi_t(X_{1:n}, c) \ge \liminf_{n\to\infty} \frac{1}{n}|R(X_{1:n})| > 0.$$

■

8. Proof of the Application to Exponential Families

In this section, we prove Theorem 11. First, we need a few supporting results. Given $y_1, \dots, y_n \in \mathbb{R}^\ell$ (for some $\ell > 0$), $\beta \in (0, 1]$, and $U \subset \mathbb{R}^\ell$, define

$$\mathcal{I}_\beta(y_{1:n}, U) = \prod_{\substack{A \subset \{1,\dots,n\}: \\ |A| \ge \beta n}} I\Big(\frac{1}{|A|}\sum_{j \in A} y_j \in U\Big), \tag{12}$$

where as usual, $I(E)$ is 1 if E is true, and 0 otherwise.

Lemma 14 (Capture lemma) *Let $V \subset \mathbb{R}^k$ be open and convex. Let Q be a probability measure on $\mathbb{R}^k$ such that:*

(1) *$\mathbb{E}|Y| < \infty$ when $Y \sim Q$,*

(2) *$Q(\bar{V}) = 1$, and*

(3) *$Q(L) = 0$ for any hyperplane L that does not intersect V.*

If $Y_1, Y_2, \dots \overset{\text{iid}}{\sim} Q$, then for any $\beta \in (0, 1]$ there exists $U \subset V$ compact such that $\mathcal{I}_\beta(Y_{1:n}, U) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$.

Proof The proof is rather long, but not terribly difficult. See Appendix D. ■

Proposition 15 *Let $Z_1, Z_2, \dots \in \mathbb{R}^k$ be i.i.d.. If $\beta \in (0, 1]$ and $U \subset \mathbb{R}^k$ such that $\mathbb{P}(Z_j \notin U) < \beta/2$, then $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$, where $Y_j = I(Z_j \in U)$.*

Proof By the law of large numbers, $\frac{1}{n}\sum_{j=1}^n I(Z_j \notin U) \xrightarrow{\text{a.s.}} \mathbb{P}(Z_j \notin U) < \beta/2$. Hence, with probability 1, for all n sufficiently large, $\frac{1}{n}\sum_{j=1}^n I(Z_j \notin U) \leq \beta/2$ holds. When it holds, we have that for any $A \subset \{1,\dots,n\}$ such that $|A| \geq \beta n$,

$$\frac{1}{|A|}\sum_{j\in A} I(Z_j \in U) = 1 - \frac{1}{|A|}\sum_{j\in A} I(Z_j \notin U) \geq 1 - \frac{1}{\beta n}\sum_{j=1}^n I(Z_j \notin U) \geq 1/2,$$

i.e., when it holds, we have $\mathcal{I}_\beta(Y_{1:n},[\frac{1}{2},1]) = 1$. Hence, $\mathcal{I}_\beta(Y_{1:n},[\frac{1}{2},1]) \xrightarrow{\text{a.s.}} 1$. ■

Given a well-behaved exponential family with conjugate prior, define

$$\mu_{x_A} = \frac{\xi + \sum_{j\in A} s(x_j)}{\nu + |A|} \tag{13}$$

(cf. Equation 14), where $x_A = (x_j : j \in A)$, $x_j \in \mathcal{X}$. In particular, $\mu_x = (\xi + s(x))/(\nu+1)$ for $x \in \mathcal{X}$.

Proposition 16 *Consider a well-behaved exponential family with conjugate prior. Let P be a probability measure on $\mathcal{X}$ such that $\mathbb{P}(s(X) \in \overline{\mathcal{M}}) = 1$ when $X \sim P$. Let $X_1, X_2, \dots \overset{\text{iid}}{\sim} P$. Then for any $\beta \in (0,1]$ there exists $U \subset \mathcal{M}$ compact such that $\mathcal{I}_\beta(Y_{1:n},[\frac{1}{2},1]) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$, where $Y_j = I(\mu_{X_j} \in U)$.*

Proof Since $\mathcal{M}$ is open and convex, then for any $y \in \overline{\mathcal{M}}$, $z \in \mathcal{M}$, and $\rho \in (0,1)$, we have $\rho y + (1-\rho)z \in \mathcal{M}$ (by e.g., Rockafellar, 1970, 6.1). Taking $z = \xi/\nu$ and $\rho = 1/(\nu+1)$, this implies that the set $U_0 = \{(\xi + y)/(\nu+1) : y \in \overline{\mathcal{M}}\}$ is contained in $\mathcal{M}$. Note that U_0 is closed and $\mathbb{P}(\mu_X \in U_0) = \mathbb{P}(s(X) \in \overline{\mathcal{M}}) = 1$. Let $\beta \in (0,1]$, and choose $r \in (0,\infty)$ such that $\mathbb{P}(|\mu_X| > r) < \beta/2$. Letting $U = \{y \in U_0 : |y| \leq r\}$, we have that $U \subset \mathcal{M}$, and U is compact. Further, $\mathbb{P}(\mu_X \notin U) < \beta/2$, so by applying Proposition 15 with $Z_j = \mu_{X_j}$, we have $\mathcal{I}_\beta(Y_{1:n},[\frac{1}{2},1]) \xrightarrow{\text{a.s.}} 1$. ■

Proposition 17 (Splitting inequality) *Consider a well-behaved exponential family with conjugate prior. For any $U \subset \mathcal{M}$ compact there exists $C \in (0,\infty)$ such that we have the following:*

For any $n \in \{1,2,\dots\}$, if $A \subset \{1,\dots,n\}$ and $B = \{1,\dots,n\} \smallsetminus A$ are nonempty, and $x_1,\dots,x_n \in \mathcal{X}$ satisfy $\frac{1}{|A|}\sum_{j\in A} s(x_j) \in U$ and $\mu_{x_B} \in U$, then

$$\frac{m(x_{1:n})}{m(x_A)m(x_B)} \leq C\left(\frac{ab}{\nu+n}\right)^{k/2}$$

where $a = \nu + |A|$ and $b = \nu + |B|$. (Recall that k is the dimension of $s : \mathcal{X} \to \mathbb{R}^k$.)

Proof See Appendix B. ■

Lemma 18 *Consider a well-behaved exponential family with conjugate prior, and the resulting collection of single-cluster marginals $m(\cdot)$. Let P be a probability measure on $\mathcal{X}$ satisfying Condition 10 (for the s and $\mathcal{M}$ from the exponential family under consideration), and let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. Then for any $\beta \in (0,1]$ there exists $c \in (0,\infty)$ such that with probability 1, for all n sufficiently large, the following event holds: for every subset $J \subset \{1,\ldots,n\}$ such that $|J| \geq \beta n$, there exists $K \subset J$ such that $|K| \geq \frac{1}{2}|J|$ and for any $j \in K$,*

$$m(X_J) \leq c\, m(X_{J \smallsetminus j})\, m(X_j).$$

Proof Let $\beta \in (0,1]$. Since $\mathcal{M}$ is open and convex, and Condition 10 holds by assumption, then by Lemma 14 (with $V = \mathcal{M}$) there exists $U_1 \subset \mathcal{M}$ compact such that $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$, where $s(X_{1:n}) = (s(X_1), \ldots, s(X_n))$. By Proposition 16 above, there exists $U_2 \subset \mathcal{M}$ compact such that $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$, where $Y_j = I(\mu_{X_j} \in U_2)$. Hence,

$$\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1)\, \mathcal{I}_\beta(Y_{1:n}, [\tfrac{1}{2}, 1]) \xrightarrow[n\to\infty]{\text{a.s.}} 1.$$

Choose $C \in (0,\infty)$ according to Proposition 17 applied to $U := U_1 \cup U_2$. We will prove the result with $c = (\nu+1)^{k/2} C$. (Recall that k is the dimension of $s : \mathcal{X} \to \mathbb{R}^k$.)

Let n large enough that $\beta n \geq 2$, and suppose that $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) = 1$ and $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) = 1$. Let $J \subset \{1,\ldots,n\}$ such that $|J| \geq \beta n$. Then for any $j \in J$,

$$\frac{1}{|J \smallsetminus j|} \sum_{i \in J \smallsetminus j} s(X_i) \in U_1 \subset U$$

since $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) = 1$ and $|J \smallsetminus j| \geq |J|/2 \geq (\beta/2)n$. Hence, for any $j \in K$, where $K = \{j \in J : \mu_{X_j} \in U\}$, we have

$$\frac{m(X_J)}{m(X_{J\smallsetminus j})\, m(X_j)} \leq C\, \Big(\frac{(\nu + |J| - 1)(\nu+1)}{\nu + |J|}\Big)^{k/2} \leq C\,(\nu+1)^{k/2} = c$$

by our choice of C above, and

$$\frac{|K|}{|J|} \geq \frac{1}{|J|} \sum_{j \in J} I(\mu_{X_j} \in U_2) = \frac{1}{|J|} \sum_{j \in J} Y_j \geq 1/2$$

since $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) = 1$ and $|J| \geq \beta n$. ■

Proof of Theorem 11 Let $t \in \{1,2,\ldots\}$ and choose c according to Lemma 18 with $\beta = 1/t$. We will show that for any $n > t$, if the event of Lemma 18 holds, then $\varphi_t(X_{1:n}, c) \geq 1/(2t)$. Since with probability 1, this event holds for all n sufficiently large, it will follow that with probability 1, $\liminf_n \varphi_t(X_{1:n}, c) \geq 1/(2t) > 0$.

So, let $n > t$ and $x_1, \ldots, x_n \in \mathcal{X}$, and assume the event of Lemma 18 holds. Let $A \in \mathcal{A}_t(n)$. There is at least one part A_ℓ such that $|A_\ell| \geq n/t = \beta n$. Then, by assumption there exists $K_A \subset A_\ell$ such that $|K_A| \geq \frac{1}{2}|A_\ell|$ and for any $j \in K_A$, $m(x_{A_\ell}) \leq c\, m(x_{A_\ell \smallsetminus j})\, m(x_j)$. Thus, $K_A \subset S_A(x_{1:n}, c)$, hence $|S_A(x_{1:n}, c)| \geq |K_A| \geq \frac{1}{2}|A_\ell| \geq n/(2t)$. Since $A \in \mathcal{A}_t(n)$ was arbitrary, $\varphi_t(x_{1:n}, c) \geq 1/(2t)$. ■

Acknowledgments

We would like to thank Stu Geman for raising this question. Special thanks to the action editor, Yee Whye Teh, and the two anonymous referees, for several helpful comments and suggestions that substantially improved the organization and content of the paper. Thanks also to Steve MacEachern, Erik Sudderth, Mike Hughes, Tamara Broderick, Annalisa Cerquetti, Dan Klein, and Dahlia Nadkarni for helpful conversations. This work was supported in part by NSF grant DMS-1007593 and DARPA contract FA8650-11-1-715.

Appendix A. Exponential Family Properties

We note some well-known properties of exponential families satisfying Condition 8. For a general reference on this material, see Hoffmann-Jørgensen (1994). Let $S_\lambda(s) = \operatorname{support}(\lambda s^{-1})$, that is,

$$S_\lambda(s) = \big\{z \in \mathbb{R}^k : \lambda(s^{-1}(U)) \neq 0 \text{ for every neighborhood } U \text{ of } z\big\}.$$

Let $C_\lambda(s)$ be the closed convex hull of $S_\lambda(s)$ (that is, the intersection of all closed convex sets containing it). Given $U \subset \mathbb{R}^k$, let U° denote its interior. Given a (sufficiently smooth) function $f : \mathbb{R}^k \to \mathbb{R}$, we use f' to denote its gradient, that is, $f'(x)_i = \frac{\partial f}{\partial x_i}(x)$, and $f''(x)$ to denote its Hessian matrix, that is, $f''(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$.

Proposition 19 *If Condition 8 is satisfied, then:*

(1) κ *is* C^∞ *smooth and strictly convex on* Θ*,*

(2) $\kappa'(\theta) = \mathbb{E}s(X)$ *and* $\kappa''(\theta) = \operatorname{Cov} s(X)$ *when* $\theta \in \Theta$ *and* $X \sim P_\theta$*,*

(3) $\kappa''(\theta)$ *is symmetric positive definite for all* $\theta \in \Theta$*,*

(4) $\kappa' : \Theta \to \mathcal{M}$ *is a* C^∞ *smooth bijection,*

(5) $\kappa'^{-1} : \mathcal{M} \to \Theta$ *is* C^∞ *smooth,*

(6) Θ *is open and convex,*

(7) $\mathcal{M}$ *is open and convex,*

(8) $\mathcal{M} = C_\lambda(s)^\circ$ *and* $\overline{\mathcal{M}} = C_\lambda(s)$*, and*

(9) $\kappa'^{-1}(\mu) = \operatorname{argmax}_{\theta\in\Theta}(\theta^{\mathsf{T}}\mu - \kappa(\theta))$ *for all* $\mu \in \mathcal{M}$*. The maximizing* $\theta \in \Theta$ *always exists and is unique.*

Proof These properties are all well-known. Let us abbreviate Hoffmann-Jørgensen (1994) as HJ. For (1), see HJ 8.36(1) and HJ 12.7.5. For (6),(2),(3), and (4), see HJ 8.36, 8.36.2-3, 12.7(2), and 12.7.11, respectively. Item (5) and openness in (7) follow, using the inverse function theorem (Knapp, 2005, 3.21). Item (8) and convexity in (7) follow, using HJ 8.36.15 and Rockafellar (1970) 6.2-3. Item (9) follows from HJ 8.36.15 and item (4). ■

Given an exponential family with conjugate prior as in Section 5.2, the joint density of $x_1, \dots, x_n \in \mathcal{X}$ and $\theta \in \mathbb{R}^k$ is

$$\begin{aligned} &p_\theta(x_1)\cdots p_\theta(x_n)\pi_{\xi,\nu}(\theta) \\ &= \exp\Big((\nu+n)\big(\theta^{\mathrm{T}}\mu_{x_{1:n}} - \kappa(\theta)\big)\Big)\exp(-\psi(\xi,\nu))\, I(\theta\in\Theta) \end{aligned} \tag{14}$$

where $\mu_{x_{1:n}} = (\xi + \sum_{j=1}^n s(x_j))/(\nu+n)$. The marginal density, defined as in Equation 5, is

$$m(x_{1:n}) = \exp\Big(\psi\big(\xi + \textstyle\sum s(x_j),\, \nu+n\big) - \psi(\xi,\nu)\Big) \tag{15}$$

when this quantity is well-defined.

Proposition 20 *If Condition 8 is satisfied, then:*

(1) *$\psi(\xi,\nu)$ is finite and C^∞ smooth on Ξ,*

(2) *if $s(x_1),\dots,s(x_n) \in S_\lambda(s)$ and $(\xi,\nu)\in\Xi$, then $(\xi+\sum s(x_j),\, \nu+n)\in\Xi$,*

(3) *$\{\pi_{\xi,\nu} : (\xi,\nu)\in\Xi\}$ is a conjugate family for $\{p_\theta : \theta\in\Theta\}$, and*

(4) *if $s : \mathcal{X}\to\mathbb{R}^k$ is continuous, $(\xi,\nu)\in\Xi$, and $\lambda(U)\neq 0$ for any nonempty $U\subset\mathcal{X}$ that is open in $\mathcal{X}$, then $m(x_{1:n}) < \infty$ for any $x_1,\dots,x_n\in\mathcal{X}$.*

Proof (1) For finiteness, see Diaconis and Ylvisaker (1979), Theorem 1. Smoothness holds for the same reason that κ is smooth; see Hoffmann-Jørgensen (1994, 8.36(1)). (Note that Ξ is open in $\mathbb{R}^{k+1}$, since $\mathcal{M}$ is open in $\mathbb{R}^k$.)

(2) Since $C_\lambda(s)$ is convex, $\frac{1}{n}\sum s(x_j) \in C_\lambda(s)$. Since $C_\lambda(s) = \bar{\mathcal{M}}$ and $\mathcal{M}$ is open and convex by 19(7) and (8), then $(\xi+\sum s(x_j))/(\nu+n)\in\mathcal{M}$, as a (strict) convex combination of $\frac{1}{n}\sum s(x_j)\in\bar{\mathcal{M}}$ and $\xi/\nu\in\mathcal{M}$ (Rockafellar, 1970, 6.1).

(3) Let $(\xi,\nu)\in\Xi$, $\theta\in\Theta$. If $X_1,\dots,X_n \overset{\text{iid}}{\sim} P_\theta$ then $s(X_1),\dots,s(X_n)\in S_\lambda(s)$ almost surely, and thus $(\xi+\sum s(X_j),\, \nu+n)\in\Xi$ (a.s.) by (2). By Equations 14 and 15, the posterior is $\pi_{\xi+\sum s(X_j),\nu+n}$.

(4) The assumptions imply $\{s(x) : x\in\mathcal{X}\}\subset S_\lambda(s)$, and therefore, for any $x_1,\dots,x_n\in\mathcal{X}$, we have $(\xi+\sum s(x_j),\, \nu+n)\in\Xi$ by (2). Thus, by (1) and Equation 15, $m(x_{1:n})<\infty$. ■

It is worth mentioning that while $\Xi\subset\big\{(\xi,\nu)\in\mathbb{R}^{k+1} : \psi(\xi,\nu)<\infty\big\}$, it may be a strict subset—often, Ξ is not quite the full set of parameters on which $\pi_{\xi,\nu}$ can be defined.

Proof of Proposition 12 (1) For any $\theta \in \Theta$ and any $j \in \{1, \ldots, k\}$,

$$\int_{\mathcal{X}} s_j(x)^2 p_\theta(x)\, d\lambda(x) = \exp(-\kappa(\theta)) \frac{\partial^2}{\partial \theta_j^2} \int_{\mathcal{X}} \exp(\theta^{\mathtt{T}} s(x))\, d\lambda(x) < \infty$$

(Hoffmann-Jørgensen, 1994, 8.36.1). Since P has density $f = \sum \pi_i p_{\theta(i)}$ with respect to λ, then

$$\mathbb{E} s_j(X)^2 = \int_{\mathcal{X}} s_j(x)^2 f(x)\, d\lambda(x) = \sum_{i=1}^{t} \pi_i \int_{\mathcal{X}} s_j(x)^2 p_{\theta(i)}(x)\, d\lambda(x) < \infty,$$

and hence

$$(\mathbb{E}|s(X)|)^2 \leq \mathbb{E}|s(X)|^2 = \mathbb{E}s_1(X)^2 + \cdots + \mathbb{E}s_k(X)^2 < \infty.$$

(2) Note that $S_P(s) \subset S_\lambda(s)$ (in fact, they are equal since P_θ and λ are mutually absolutely continuous for any $\theta \in \Theta$), and therefore

$$S_P(s) \subset S_\lambda(s) \subset C_\lambda(s) = \overline{\mathcal{M}}$$

by Proposition 19(8). Hence,

$$\mathbb{P}(s(X) \in \overline{\mathcal{M}}) \geq \mathbb{P}(s(X) \in S_P(s)) = Ps^{-1}(\text{support}(Ps^{-1})) = 1.$$

(3) Suppose λ is absolutely continuous with respect to Lebesgue measure, $\mathcal{X}$ is open and connected, and s is real analytic. Let $L \subset \mathbb{R}^k$ be a hyperplane, and write $L = \{z \in \mathbb{R}^k : z^{\mathtt{T}} y = b\}$ where $y \in \mathbb{R}^k \smallsetminus \{0\}$, $b \in \mathbb{R}$. Define $g : \mathcal{X} \to \mathbb{R}$ by $g(x) = s(x)^{\mathtt{T}} y - b$. Then g is real analytic on $\mathcal{X}$, since a finite sum of real analytic functions is real analytic. Since $\mathcal{X}$ is connected, it follows that either g is identically zero, or the set $V = \{x \in \mathcal{X} : g(x) = 0\}$ has Lebesgue measure zero (Krantz, 1992). Now, g cannot be identically zero, since for any $\theta \in \Theta$, letting $Z \sim P_\theta$, we have

$$0 < y^{\mathtt{T}} \kappa''(\theta) y = y^{\mathtt{T}} (\operatorname{Cov} s(Z)) y = \operatorname{Var}(y^{\mathtt{T}} s(Z)) = \operatorname{Var} g(Z)$$

by Proposition 19(2) and (3). Consequently, V must have Lebesgue measure zero. Hence, $P(V) = 0$, since P is absolutely continuous with respect to λ, and thus, with respect to Lebesgue measure. Therefore,

$$\mathbb{P}(s(X) \in L) = \mathbb{P}(g(X) = 0) = P(V) = 0.$$

■

Appendix B. Marginal Inequalities

In this section, we prove Proposition 17, which was used in the key lemma for the exponential family case (Lemma 18).

Consider a well-behaved exponential family with conjugate prior (as in Definition 9). The proof uses some simple bounds on the Laplace approximation (see Appendix C) to obtain inequalities involving the marginal density $m(x_{1:n})$ (cf. Equations 5 and 15) of

$x_{1:n} = (x_1, \dots, x_n)$, where $x_j \in \mathcal{X}$. Of course, it is commonplace to apply the Laplace approximation to $m(X_{1:n})$ when $X_1, \dots, X_n$ are i.i.d. random variables and n is sufficiently large. In contrast, our application of it is considerably more subtle. For our purposes, it is necessary to show that for every n, even without assuming i.i.d. data, the approximation is good enough as long as the sufficient statistics are not too extreme.

We make extensive use of the exponential family properties in Appendix A, often without mention. We use f' to denote the gradient and f'' to denote the Hessian of a (sufficiently smooth) function $f : \mathbb{R}^k \to \mathbb{R}$. For $\mu \in \mathcal{M}$, define

$$\begin{aligned} f_\mu(\theta) &= \theta^{\mathtt{T}}\mu - \kappa(\theta), \\ \mathcal{L}(\mu) &= \sup_{\theta \in \Theta} \big(\theta^{\mathtt{T}}\mu - \kappa(\theta)\big), \\ \theta_\mu &= \operatorname*{argmax}_{\theta \in \Theta} \big(\theta^{\mathtt{T}}\mu - \kappa(\theta)\big), \end{aligned}$$

and note that $\theta_\mu = \kappa'^{-1}(\mu)$ (Proposition 19). $\mathcal{L}$ is known as the Legendre transform of κ. Note that $\mathcal{L}(\mu) = f_\mu(\theta_\mu)$, and $\mathcal{L}$ is C^∞ smooth on $\mathcal{M}$ (since $\mathcal{L}(\mu) = \theta_\mu^{\mathtt{T}}\mu - \kappa(\theta_\mu)$, $\theta_\mu = \kappa'^{-1}(\mu)$, and both κ and κ'^{-1} are C^∞ smooth). As in Equation 13, define

$$\mu_{x_{1:n}} = \frac{\xi + \sum_{j=1}^n s(x_j)}{\nu + n} \tag{16}$$

and given $x_{1:n}$ such that $\mu_{x_{1:n}} \in \mathcal{M}$, define

$$\widetilde{m}(x_{1:n}) = (\nu + n)^{-k/2} \exp\big((\nu + n)\, \mathcal{L}(\mu_{x_{1:n}})\big),$$

where k is the dimension of the sufficient statistics function $s : \mathcal{X} \to \mathbb{R}^k$. Proposition 21 below provides uniform bounds on $m(x_{1:n})/\widetilde{m}(x_{1:n})$. Here, $\widetilde{m}(x_{1:n})$ is only intended to approximate $m(x_{1:n})$ up to a multiplicative constant—a better approximation could always be obtained via the usual asymptotic form of the Laplace approximation.

Proposition 21 *Consider a well-behaved exponential family with conjugate prior. For any $U \subset \mathcal{M}$ compact, there exist $C_1, C_2 \in (0, \infty)$ such that for any $n \in \{1, 2, \dots\}$ and any $x_1, \dots, x_n \in \mathcal{X}$ satisfying $\mu_{x_{1:n}} \in U$, we have*

$$C_1 \le \frac{m(x_{1:n})}{\widetilde{m}(x_{1:n})} \le C_2.$$

Proof Assume $U \neq \varnothing$, since otherwise the result is trivial. Let

$$V = \kappa'^{-1}(U) = \{\theta_\mu : \mu \in U\}.$$

It is straightforward to show that there exists $\varepsilon \in (0, 1)$ such that $V_\varepsilon \subset \Theta$ where

$$V_\varepsilon = \{\theta \in \mathbb{R}^k : d(\theta, V) \le \varepsilon\}.$$

(Here, $d(\theta, V) = \inf_{\theta' \in V} |\theta - \theta'|$.) Note that V_ε is compact, since κ'^{-1} is continuous. Given a symmetric matrix A, define $\lambda_*(A)$ and $\lambda^*(A)$ to be the minimal and maximal eigenvalues, respectively, and recall that λ_*, λ^* are continuous functions of the entries of A. Letting

$$\alpha = \min_{\theta \in V_\varepsilon} \lambda_*(\kappa''(\theta)) \quad \text{and} \quad \beta = \max_{\theta \in V_\varepsilon} \lambda^*(\kappa''(\theta)),$$

we have $0 < \alpha \leq \beta < \infty$ since V_ε is compact and $\lambda_*(\kappa''(\cdot))$, $\lambda^*(\kappa''(\cdot))$ are continuous and positive on Θ. Letting

$$\gamma = \sup_{\mu \in U} e^{-f_\mu(\theta_\mu)} \int_\Theta \exp(f_\mu(\theta)) d\theta = \sup_{\mu \in U} e^{-\mathcal{L}(\mu)} e^{\psi(\mu,1)}$$

we have $0 < \gamma < \infty$ since U is compact, and both $\mathcal{L}$ (as noted above) and $\psi(\mu, 1)$ (by Proposition 20) are continuous on $\mathcal{M}$. Define

$$h(\mu, \theta) = f_\mu(\theta_\mu) - f_\mu(\theta) = \mathcal{L}(\mu) - \theta^{\mathtt{T}}\mu + \kappa(\theta)$$

for $\mu \in \mathcal{M}$, $\theta \in \Theta$. For any $\mu \in \mathcal{M}$, we have that $h(\mu, \theta) > 0$ whenever $\theta \in \Theta \smallsetminus \{\theta_\mu\}$, and that $h(\mu, \theta)$ is strictly convex in θ. Letting $B_\varepsilon(\theta_\mu) = \{\theta \in \mathbb{R}^k : |\theta - \theta_\mu| \leq \varepsilon\}$, it follows that

$$\delta := \inf_{\mu \in U} \inf_{\theta \in \Theta \smallsetminus B_\varepsilon(\theta_\mu)} h(\mu, \theta) = \inf_{\mu \in U} \inf_{u \in \mathbb{R}^k : |u| = 1} h(\mu, \theta_\mu + \varepsilon u)$$

is positive, as the minimum of a positive continuous function on a compact set.

Now, applying the Laplace approximation bounds in Corollary 24 with $\alpha, \beta, \gamma, \delta, \varepsilon$ as just defined, we obtain $c_1, c_2 \in (0, \infty)$ such that for any $\mu \in U$ we have (taking $E = \Theta$, $f = -f_\mu$, $x_0 = \theta_\mu$, $A = \alpha I_{k \times k}$, $B = \beta I_{k \times k}$)

$$c_1 \leq \frac{\int_\Theta \exp(t f_\mu(\theta)) d\theta}{t^{-k/2} \exp(t f_\mu(\theta_\mu))} \leq c_2$$

for any $t \geq 1$. We prove the result with $C_i = c_i\, e^{-\psi(\xi,\nu)}$ for $i = 1, 2$.

Let $n \in \{1, 2, \dots\}$ and $x_1, \dots, x_n \in \mathcal{X}$ such that $\mu_{x_{1:n}} \in U$. Choose $t = \nu + n$. By integrating Equation 14, we have

$$m(x_{1:n}) = e^{-\psi(\xi,\nu)} \int_\Theta \exp\left(t f_{\mu_{x_{1:n}}}(\theta)\right) d\theta,$$

and meanwhile,

$$\widetilde{m}(x_{1:n}) = t^{-k/2} \exp\left(t f_{\mu_{x_{1:n}}}(\theta_{\mu_{x_{1:n}}})\right).$$

Thus, combining the preceding three displayed equations,

$$0 < C_1 = c_1 e^{-\psi(\xi,\nu)} \leq \frac{m(x_{1:n})}{\widetilde{m}(x_{1:n})} \leq c_2 e^{-\psi(\xi,\nu)} = C_2 < \infty.$$

■

Proof of Proposition 17 Let U' be the convex hull of $U \cup \{\xi/\nu\}$. Then U' is compact (as the convex hull of a compact set in $\mathbb{R}^k$) and $U' \subset \mathcal{M}$ (since $U \cup \{\xi/\nu\} \subset \mathcal{M}$ and $\mathcal{M}$ is convex). We show that the result holds with $C = C_2 \exp(C_0)/C_1^2$, where $C_1, C_2 \in (0, \infty)$ are obtained by applying Proposition 21 to U', and

$$C_0 = \nu \sup_{y \in U'} |(\xi/\nu - y)^{\mathtt{T}} \mathcal{L}'(y)| + \nu \sup_{y \in U'} |\mathcal{L}(y)| < \infty. \tag{17}$$

Since $\mathcal{L}$ is convex (being a Legendre transform) and smooth, then for any $y, z \in \mathcal{M}$ we have

$$\inf_{\rho\in(0,1)} \frac{1}{\rho}\big(\mathcal{L}(y+\rho(z-y)) - \mathcal{L}(y)\big) = (z-y)^{\mathtt{T}}\mathcal{L}'(y)$$

(by e.g., Rockafellar, 1970, 23.1) and therefore for any $\rho \in (0,1)$,

$$\mathcal{L}(y) \leq \mathcal{L}((1-\rho)y+\rho z) - \rho(z-y)^{\mathtt{T}}\mathcal{L}'(y). \tag{18}$$

Choosing $y = \mu_{x_{1:n}}$, $z = \xi/\nu$, and $\rho = \nu/(n+2\nu)$, we have

$$(1-\rho)y + \rho z = \frac{2\xi + \sum_{j=1}^n s(x_j)}{2\nu+n} = \frac{a\mu_{x_A} + b\mu_{x_B}}{a+b}. \tag{19}$$

Note that $\mu_{x_A}, \mu_{x_B}, \mu_{x_{1:n}} \in U'$, by taking various convex combinations of ξ/ν, $\frac{1}{|A|}\sum_{j\in A} s(x_j)$, $\mu_{x_B} \in U'$. Thus,

$$\begin{aligned}(\nu+n)\mathcal{L}(\mu_{x_{1:n}}) &= (a+b)\mathcal{L}(y) - \nu\mathcal{L}(y)\\ &\overset{\text{(a)}}{\leq} (a+b)\mathcal{L}((1-\rho)y+\rho z) - (a+b)\rho(z-y)^{\mathtt{T}}\mathcal{L}'(y) - \nu\mathcal{L}(y)\\ &\overset{\text{(b)}}{\leq} (a+b)\mathcal{L}\Big(\frac{a\mu_{x_A}+b\mu_{x_B}}{a+b}\Big) + C_0\\ &\overset{\text{(c)}}{\leq} a\mathcal{L}(\mu_{x_A}) + b\mathcal{L}(\mu_{x_B}) + C_0,\end{aligned}$$

where (a) is by Equation 18, (b) is by Equations 17 and 19, and (c) is by the convexity of $\mathcal{L}$. Hence, $(\nu+n)^{k/2}\widetilde{m}(x_{1:n}) \leq (ab)^{k/2}\widetilde{m}(x_A)\widetilde{m}(x_B)\exp(C_0)$, so by our choice of C_1 and C_2,

$$\frac{m(x_{1:n})}{m(x_A)m(x_B)} \leq \frac{C_2\widetilde{m}(x_{1:n})}{C_1^2\widetilde{m}(x_A)\widetilde{m}(x_B)} \leq \frac{C_2\exp(C_0)}{C_1^2}\Big(\frac{ab}{n+\nu}\Big)^{k/2}.$$

■

Appendix C. Bounds on the Laplace Approximation

Our proof uses the following simple bounds on the Laplace approximation. These bounds are not fundamentally new, but the precise formulation we require does not seem to appear in the literature, so we have included it for the reader's convenience. Lemma 22 is simply a multivariate version of the bounds given by De Bruijn (1970), and Corollary 24 is a straightforward consequence, putting the lemma in a form most convenient for our purposes.

Given symmetric matrices A and B, let us write $A \trianglelefteq B$ to mean that $B - A$ is positive semidefinite. Given $A \in \mathbb{R}^{k\times k}$ symmetric positive definite and $\varepsilon, t \in (0,\infty)$, define

$$C(t,\varepsilon,A) = \mathbb{P}(|A^{-1/2}Z| \leq \varepsilon\sqrt{t})$$

where $Z \sim \text{Normal}(0, I_{k\times k})$. Note that $C(t,\varepsilon,A) \to 1$ as $t\to\infty$. Let $B_\varepsilon(x_0) = \{x \in \mathbb{R}^k : |x - x_0| \leq \varepsilon\}$ denote the closed ball of radius $\varepsilon > 0$ at $x_0 \in \mathbb{R}^k$.

Lemma 22 *Let $E \subset \mathbb{R}^k$ be open. Let $f : E \to \mathbb{R}$ be C^2 smooth with $f'(x_0) = 0$ for some $x_0 \in E$. Define*

$$g(t) = \int_E \exp(-tf(x))\, dx$$

for $t \in (0, \infty)$. Suppose $\varepsilon \in (0, \infty)$ such that $B_\varepsilon(x_0) \subset E$, $0 < \delta \le \inf\{f(x) - f(x_0) : x \in E \smallsetminus B_\varepsilon(x_0)\}$, and A, B are symmetric positive definite matrices such that $A \trianglelefteq f''(x) \trianglelefteq B$ for all $x \in B_\varepsilon(x_0)$. Then for any $0 < s \le t$ we have

$$\frac{C(t, \varepsilon, B)}{|B|^{1/2}} \le \frac{g(t)}{(2\pi/t)^{k/2} e^{-tf(x_0)}} \le \frac{C(t, \varepsilon, A)}{|A|^{1/2}} + \Big(\frac{t}{2\pi}\Big)^{k/2} e^{-(t-s)\delta} e^{sf(x_0)} g(s)$$

where $|A| = |\det A|$.

Remark 23 *In particular, these assumptions imply f is strictly convex on $B_\varepsilon(x_0)$ with unique global minimum at x_0. Note that the upper bound is trivial unless $g(s) < \infty$.*

Proof By Taylor's theorem, for any $x \in B_\varepsilon(x_0)$ there exists z_x on the line between x_0 and x such that, letting $y = x - x_0$,

$$f(x) = f(x_0) + y^{\mathtt{T}} f'(x_0) + \tfrac{1}{2} y^{\mathtt{T}} f''(z_x) y = f(x_0) + \tfrac{1}{2} y^{\mathtt{T}} f''(z_x) y.$$

Since $z_x \in B_\varepsilon(x_0)$, and thus $A \trianglelefteq f''(z_x) \trianglelefteq B$,

$$\tfrac{1}{2} y^{\mathtt{T}} A y \le f(x) - f(x_0) \le \tfrac{1}{2} y^{\mathtt{T}} B y.$$

Hence,

$$\begin{aligned} e^{tf(x_0)} \int_{B_\varepsilon(x_0)} \exp(-tf(x))\, dx &\le \int_{B_\varepsilon(x_0)} \exp(-\tfrac{1}{2}(x - x_0)^{\mathtt{T}} (tA)(x - x_0))\, dx \\ &= (2\pi)^{k/2} |(tA)^{-1}|^{1/2}\ \mathbb{P}\big(|(tA)^{-1/2} Z| \le \varepsilon\big). \end{aligned}$$

Along with a similar argument for the lower bound, this implies

$$\Big(\frac{2\pi}{t}\Big)^{k/2} \frac{C(t, \varepsilon, B)}{|B|^{1/2}} \le e^{tf(x_0)} \int_{B_\varepsilon(x_0)} \exp(-tf(x))\, dx \le \Big(\frac{2\pi}{t}\Big)^{k/2} \frac{C(t, \varepsilon, A)}{|A|^{1/2}}.$$

Considering the rest of the integral, outside of $B_\varepsilon(x_0)$, we have

$$0 \le \int_{E \smallsetminus B_\varepsilon(x_0)} \exp(-tf(x))\, dx \le \exp\big(-(t - s)(f(x_0) + \delta)\big)\, g(s).$$

Combining the preceding four inequalities yields the result. ■

The following corollary tailors the lemma to our purposes. Given a symmetric positive definite matrix $A \in \mathbb{R}^{k \times k}$, let $\lambda_*(A)$ and $\lambda^*(A)$ be the minimal and maximal eigenvalues, respectively. By diagonalizing A, it is easy to check that $\lambda_*(A) I_{k \times k} \trianglelefteq A \trianglelefteq \lambda^*(A) I_{k \times k}$ and $\lambda_*(A)^k \le |A| \le \lambda^*(A)^k$.

Corollary 24 *For any $\alpha, \beta, \gamma, \delta, \varepsilon \in (0, \infty)$ there exist $c_1 = c_1(\beta, \varepsilon) \in (0, \infty)$ and $c_2 = c_2(\alpha, \gamma, \delta) \in (0, \infty)$ such that if E, f, x_0, A, B satisfy all the conditions of Lemma 22 (for this choice of δ, ε) and additionally, $\alpha \leq \lambda_*(A)$, $\beta \geq \lambda^*(B)$, and $\gamma \geq e^{f(x_0)} g(1)$, then*

$$c_1 \leq \frac{\int_E \exp(-tf(x))\, dx}{t^{-k/2} \exp(-tf(x_0))} \leq c_2$$

for all $t \geq 1$.

Proof The first term in the upper bound of the lemma is $C(t, \varepsilon, A)/|A|^{1/2} \leq 1/\alpha^{k/2}$, and with $s = 1$ the second term is less or equal to $(t/2\pi)^{k/2} e^{-(t-1)\delta} \gamma$, which is bounded above for $t \in [1, \infty)$. For the lower bound, a straightforward calculation (using $z^{\mathtt{T}} B z \leq \lambda^*(B) z^{\mathtt{T}} z \leq \beta z^{\mathtt{T}} z$ in the exponent inside the integral) shows that $C(t, \varepsilon, B)/|B|^{1/2} \geq \mathbb{P}(|Z| \leq \varepsilon\sqrt{\beta})/\beta^{k/2}$ for $t \geq 1$. ■

Although we do not need it (and thus, we omit the proof), the following corollary gives the well-known asymptotic form of the Laplace approximation. (As usual, $g(t) \sim h(t)$ as $t \to \infty$ means that $g(t)/h(t) \to 1$.)

Corollary 25 *Let $E \subset \mathbb{R}^k$ be open. Let $f : E \to \mathbb{R}$ be C^2 smooth such that for some $x_0 \in E$ we have that $f'(x_0) = 0$, $f''(x_0)$ is positive definite, and $f(x) > f(x_0)$ for all $x \in E \smallsetminus \{x_0\}$. Suppose there exists $\varepsilon > 0$ such that $B_\varepsilon(x_0) \subset E$ and $\inf\{f(x) - f(x_0) : x \in E \smallsetminus B_\varepsilon(x_0)\}$ is positive, and suppose there is some $s > 0$ such that $\int_E e^{-sf(x)}\, dx < \infty$. Then*

$$\int_E \exp(-tf(x))\, dx \;\sim\; \Big(\frac{2\pi}{t}\Big)^{k/2} \frac{\exp(-tf(x_0))}{|f''(x_0)|^{1/2}}$$

as $t \to \infty$.

Appendix D. Capture Lemma

In this section, we prove Lemma 14, which is restated here for the reader's convenience.

The following definitions are standard. Let $\mathcal{S}$ denote the unit sphere in $\mathbb{R}^k$, that is, $\mathcal{S} = \{x \in \mathbb{R}^k : |x| = 1\}$. We say that $H \subset \mathbb{R}^k$ is a *halfspace* if $H = \{x \in \mathbb{R}^k : x^{\mathtt{T}} u \prec b\}$, where $\prec$ is either $<$ or $\leq$, for some $u \in \mathcal{S}$, $b \in \mathbb{R}$. We say that $L \subset \mathbb{R}^k$ is a *hyperplane* if $L = \{x \in \mathbb{R}^k : x^{\mathtt{T}} u = b\}$ for some $u \in \mathcal{S}$, $b \in \mathbb{R}$. Given $U \subset \mathbb{R}^k$, let ∂U denote the *boundary of* U, that is, $\partial U = \bar{U} \smallsetminus U^\circ$. So, for example, if H is a halfspace, then ∂H is a hyperplane. The following notation is also useful: given $x \in \mathbb{R}^k$, we call the set $R_x = \{ax : a > 0\}$ the *ray through* x.

We give the central part of the proof first, postponing some plausible intermediate results for the moment. Recall the definition of $\mathcal{I}_\beta(x_{1:n}, U)$ from Equation 12.

Lemma 26 (Capture lemma) *Let $V \subset \mathbb{R}^k$ be open and convex. Let P be a probability measure on $\mathbb{R}^k$ such that:*

1. *$\mathbb{E}|X| < \infty$ when $X \sim P$,*
2. *$P(\bar{V}) = 1$, and*

(3) *$P(L) = 0$ for any hyperplane L that does not intersect V.*

If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$, then for any $\beta \in (0,1]$ there exists $U \subset V$ compact such that $\mathcal{I}_\beta(X_{1:n}, U) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$.

Proof Without loss of generality, we may assume $0 \in V$ (since otherwise we can translate to make it so, obtain U, and translate back). Let $\beta \in (0,1]$. By Proposition 28 below, for each $u \in \mathcal{S}$ there is a closed halfspace H_u such that $0 \in H_u^\circ$, R_u intersects $V \cap \partial H_u$, and $\mathcal{I}_\beta(X_{1:n}, H_u) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$. By Proposition 30 below, there exist $u_1, \ldots, u_r \in \mathcal{S}$ (for some $r > 0$) such that the set $U = \bigcap_{i=1}^r H_{u_i}$ is compact and $U \subset V$. Finally,

$$\mathcal{I}_\beta(X_{1:n}, U) = \prod_{i=1}^r \mathcal{I}_\beta(X_{1:n}, H_{u_i}) \xrightarrow[n\to\infty]{\text{a.s.}} 1.$$

■

The main idea of the lemma is exhibited in the following simpler case, which we will use to prove Proposition 28.

Proposition 27 *Let $V = (-\infty, c)$, where $-\infty < c \leq \infty$. Let P be a probability measure on $\mathbb{R}$ such that:*

(1) *$\mathbb{E}|X| < \infty$ when $X \sim P$, and*

(2) *$P(V) = 1$.*

If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$, then for any $\beta \in (0,1]$ there exists $b < c$ such that $\mathcal{I}_\beta(X_{1:n}, (-\infty, b]) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$.

Proof Let $\beta \in (0,1]$. By continuity from above, there exists $a < c$ such that $\mathbb{P}(X > a) < \beta$. If $\mathbb{P}(X > a) = 0$ then the result is trivial, taking $b = a$. Suppose $\mathbb{P}(X > a) > 0$. Let b such that $\mathbb{E}(X \mid X > a) < b < c$, which is always possible, by a straightforward argument (using $\mathbb{E}|X| < \infty$ in the $c = \infty$ case). Let $B_n = B_n(X_1, \ldots, X_n) = \{i \in \{1, \ldots, n\} : X_i > a\}$. Then

$$\begin{aligned} \frac{1}{|B_n|} \sum_{i \in B_n} X_i &= \frac{1}{\frac{1}{n}|B_n|} \frac{1}{n} \sum_{i=1}^n X_i \, I(X_i > a) \\ &\xrightarrow[n\to\infty]{\text{a.s.}} \frac{\mathbb{E}(X \, I(X > a))}{\mathbb{P}(X > a)} = \mathbb{E}(X \mid X > a) < b. \end{aligned}$$

Now, fix $n \in \{1, 2, \ldots\}$, and suppose $0 < |B_n| < \beta n$ and $\frac{1}{|B_n|} \sum_{i \in B_n} X_i < b$, noting that with probability 1, this happens for all n sufficiently large. We show that this implies $\mathcal{I}_\beta(X_{1:n}, (-\infty, b]) = 1$. This will prove the result.

Let $A \subset \{1, \ldots, n\}$ such that $|A| \geq \beta n$. Let $M = \{\pi_1, \ldots, \pi_{|A|}\}$ where π is a permutation of $\{1, \ldots, n\}$ such that $X_{\pi_1} \geq \cdots \geq X_{\pi_n}$ (that is, $M \subset \{1, \ldots, n\}$ consists of the indices of

$|A|$ of the largest entries of $(X_1, \dots, X_n)$). Then $|M| = |A| \geq \beta n \geq |B_n|$, and it follows that $B_n \subset M$. Therefore,

$$\frac{1}{|A|}\sum_{i\in A} X_i \leq \frac{1}{|M|}\sum_{i\in M} X_i \leq \frac{1}{|B_n|}\sum_{i\in B_n} X_i \leq b,$$

as desired. ■

The first of the two propositions used in Lemma 26 is the following.

Proposition 28 *Let V and P satisfy the conditions of Lemma 26, and also assume $0 \in V$. If $X_1, X_2, \dots \overset{\text{iid}}{\sim} P$ then for any $\beta \in (0,1]$ and any $u \in \mathcal{S}$ there is a closed halfspace $H \subset \mathbb{R}^k$ such that*

(1) *$0 \in H^\circ$,*

(2) *R_u intersects $V \cap \partial H$, and*

(3) *$\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$.*

Proof Let $\beta \in (0,1]$ and $u \in \mathcal{S}$. Either (a) $R_u \subset V$, or (b) R_u intersects ∂V.

Case (a): Suppose $R_u \subset V$. Let $Y_i = X_i^{\mathtt{T}} u$ for $i = 1, 2, \dots$. Then $\mathbb{E}|Y_i| \leq \mathbb{E}|X_i||u| = \mathbb{E}|X_i| < \infty$, and thus, by Proposition 27 (with $c = \infty$) there exists $b \in \mathbb{R}$ such that $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b]) \xrightarrow{\text{a.s.}} 1$. Let us choose this b to be positive, which is always possible since $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b])$ is nondecreasing as a function of b. Let $H = \{x \in \mathbb{R}^k : x^{\mathtt{T}} u \leq b\}$. Then $0 \in H^\circ$, since $b > 0$, and R_u intersects $V \cap \partial H$ at bu, since $R_u \subset V$ and $bu^{\mathtt{T}} u = b$. And since $\frac{1}{|A|}\sum_{i\in A} Y_i \leq b$ if and only if $\frac{1}{|A|}\sum_{i\in A} X_i \in H$, we have $\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$.

Case (b): Suppose R_u intersects ∂V at some point $z \in \mathbb{R}^k$. Note that $z \neq 0$ since $0 \notin R_u$. Since $\bar{V}$ is convex, it has a supporting hyperplane at z, and thus, there exist $v \in \mathcal{S}$ and $c \in \mathbb{R}$ such that $G = \{x \in \mathbb{R}^k : x^{\mathtt{T}} v \leq c\}$ satisfies $\bar{V} \subset G$ and $z \in \partial G$ (Rockafellar, 1970, 11.2). Note that $c > 0$ and $V \cap \partial G = \varnothing$ since $0 \in V$ and V is open. Letting $Y_i = X_i^{\mathtt{T}} v$ for $i = 1, 2, \dots$, we have

$$\mathbb{P}(Y_i \leq c) = \mathbb{P}(X_i^{\mathtt{T}} v \leq c) = \mathbb{P}(X_i \in G) \geq \mathbb{P}(X_i \in \bar{V}) = P(\bar{V}) = 1,$$

and hence,

$$\mathbb{P}(Y_i \geq c) = \mathbb{P}(Y_i = c) = \mathbb{P}(X_i^{\mathtt{T}} v = c) = \mathbb{P}(X_i \in \partial G) = P(\partial G) = 0,$$

by our assumptions on P, since ∂G is a hyperplane that does not intersect V. Consequently, $\mathbb{P}(Y_i < c) = 1$. Also, as before, $\mathbb{E}|Y_i| < \infty$. Thus, by Proposition 27, there exists $b < c$ such that $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b]) \xrightarrow{\text{a.s.}} 1$. Since $c > 0$, we may choose this b to be positive (as before). Letting $H = \{x \in \mathbb{R}^k : x^{\mathtt{T}} v \leq b\}$, we have $\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$. Also, $0 \in H^\circ$ since $b > 0$.

Now, we must show that R_u intersects $V \cap \partial H$. First, since $z \in R_u$ means $z = au$ for some $a > 0$, and since $z \in \partial G$ means $z^{\mathtt{T}} v = c > 0$, we find that $u^{\mathtt{T}} v > 0$ and $z = cu/u^{\mathtt{T}} v$. Therefore, letting $y = bu/u^{\mathtt{T}} v$, we have $y \in R_u \cap V \cap \partial H$, since

(i) $b/u^{\mathtt{T}} v > 0$, and thus $y \in R_u$,

(ii) $y^{\mathsf{T}}v = b$, and thus $y \in \partial H$,

(iii) $0 < b/u^{\mathsf{T}}v < c/u^{\mathsf{T}}v$, and thus y is a (strict) convex combination of $0 \in V$ and $z \in \bar{V}$, hence $y \in V$ (Rockafellar, 1970, 6.1).

■

To prove Proposition 30, we need the following geometrically intuitive facts.

Proposition 29 *Let $V \subset \mathbb{R}^k$ be open and convex, with $0 \in V$. Let H be a closed halfspace such that $0 \in H^\circ$. Let $T = \{x/|x| : x \in V \cap \partial H\}$. Then*

(1) *T is open in $\mathcal{S}$,*

(2) *$T = \{u \in \mathcal{S} : R_u$ intersects $V \cap \partial H\}$, and*

(3) *if $x \in H$, $x \neq 0$, and $x/|x| \in T$, then $x \in V$.*

Proof Write $H = \{x \in \mathbb{R}^k : x^{\mathsf{T}}v \leq b\}$, with $v \in \mathcal{S}$, $b > 0$. Let $\mathcal{S}_+ = \{u \in \mathcal{S} : u^{\mathsf{T}}v > 0\}$. (1) Define $f : \partial H \to \mathcal{S}_+$ by $f(x) = x/|x|$, noting that $0 \notin \partial H$. It is easy to see that f is a homeomorphism. Since V is open in $\mathbb{R}^k$, then $V \cap \partial H$ is open in ∂H. Hence, $T = f(V \cap \partial H)$ is open in $\mathcal{S}_+$, and since $\mathcal{S}_+$ is open in $\mathcal{S}$, then T is also open in $\mathcal{S}$. Items (2) and (3) are easily checked. ■

Proposition 30 *Let $V \subset \mathbb{R}^k$ be open and convex, with $0 \in V$. If $(H_u : u \in \mathcal{S})$ is a collection of closed halfspaces such that for all $u \in \mathcal{S}$,*

(1) *$0 \in H_u^\circ$ and*

(2) *R_u intersects $V \cap \partial H_u$,*

then there exist $u_1, \ldots, u_r \in \mathcal{S}$ (for some $r > 0$) such that the set $U = \bigcap_{i=1}^r H_{u_i}$ is compact and $U \subset V$.

Proof For $u \in \mathcal{S}$, define $T_u = \{x/|x| : x \in V \cap \partial H_u\}$. By part (1) of Proposition 29, T_u is open in $\mathcal{S}$, and by part (2), $u \in T_u$, since R_u intersects $V \cap \partial H_u$. Thus, $(T_u : u \in \mathcal{S})$ is an open cover of $\mathcal{S}$. Since $\mathcal{S}$ is compact, there is a finite subcover: there exist $u_1, \ldots, u_r \in \mathcal{S}$ (for some $r > 0$) such that $\bigcup_{i=1}^r T_{u_i} \supset \mathcal{S}$, and in fact, $\bigcup_{i=1}^r T_{u_i} = \mathcal{S}$. Let $U = \bigcap_{i=1}^r H_{u_i}$. Then U is closed and convex (as an intersection of closed, convex sets). Further, $U \subset V$ since for any $x \in U$, if $x = 0$ then $x \in V$ by assumption, while if $x \neq 0$ then $x/|x| \in T_{u_i}$ for some $i \in \{1, \ldots, r\}$ and $x \in U \subset H_{u_i}$, so $x \in V$ by Proposition 29(3).

In order to show that U is compact, we just need to show it is bounded, since we already know it is closed. Suppose not, and let $x_1, x_2, \ldots \in U \smallsetminus \{0\}$ such that $|x_n| \to \infty$ as $n \to \infty$. Let $v_n = x_n/|x_n|$. Since $\mathcal{S}$ is compact, then (v_n) has a convergent subsequence such that $v_{n_i} \to u$ for some $u \in \mathcal{S}$. Then for any $a > 0$, we have $av_{n_i} \in U$ for all i sufficiently large (since av_{n_i} is a convex combination of $0 \in U$ and $|x_{n_i}|v_{n_i} = x_{n_i} \in U$ whenever $|x_{n_i}| \geq a$). Since $av_{n_i} \to au$, and U is closed, then $au \in U$. Thus, $au \in U$ for all $a > 0$, i.e., $R_u \subset U$.

But $u \in T_{u_j}$ for some $j \in \{1, \ldots, r\}$, so R_u intersects ∂H_{u_j} (by Proposition 29(2)), and thus $au \notin H_{u_j} \supset U$ for all $a > 0$ sufficiently large. This is a contradiction. Therefore, U is bounded. ■

References

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, November 1974.

R. Argiento, A. Guglielmi, and A. Pievatolo. A comparison of nonparametric priors in hierarchical mixture modelling for AFT regression. *Journal of Statistical Planning and Inference*, 139(12):3989–4005, 2009.

D. A. Berry and R. Christensen. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, pages 558–568, 1979.

A. Bhattacharya and D. B. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97(4):851–865, 2010.

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

C. A. Bush and S. N. MacEachern. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

H. Chen, P. L. Morrell, V. E. T. M. Ashworth, M. de la Cruz, and M. T. Clegg. Tracing the geographic origins of major avocado cultivars. *Journal of Heredity*, 100(1):56–65, 2009.

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

N. G. De Bruijn. *Asymptotic Methods in Analysis.* North-Holland Publishing Co., Amsterdam, 1970.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.

M. D. Escobar. *Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means.* PhD thesis, Yale University, 1988.

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

M. D. Escobar and M. West. Computing nonparametric hierarchical models. In D. Dey, P. Müller, and D. Sinha, editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 1–22. Springer-Verlag, New York, 1998.

P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21, 2004.

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

T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics*, pages 287–302. Academic Press, 1983.

E. B. Fox, E. B. Sudderth, and A. S. Willsky. Hierarchical Dirichlet processes for tracking maneuvering targets. In *10th International Conference on Information Fusion, 2007*, pages 1–8. IEEE, 2007.

S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 36–83. Cambridge University Press, 2010.

S. Ghosal and A. Van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

S. Ghosal and A. W. Van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, pages 1233–1263, 2001.

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

S. K. Ghosh and S. Ghosal. Semiparametric accelerated failure time models for censored data. *Bayesian Statistics and its Applications*, pages 213–229, 2006.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

E. G. Gonzalez and R. Zardoya. Relative role of life-history traits and historical factors in shaping genetic population structure of sardines (Sardina pilchardus). *BMC Evolutionary Biology*, 7(1):197, 2007.

P. J. Green and S. Richardson. Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, June 2001.

B. Hansen and J. Pitman. Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters*, 46(3):251–256, 2000.

J. Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(1):235–240, 1985.

J. Henna. Estimation of the number of components of finite mixtures of multivariate distributions. *Annals of the Institute of Statistical Mathematics*, 57(4):655–664, 2005.

J. Hoffmann-Jørgensen. *Probability With a View Towards Statistics*, volume 2. Chapman & Hall, 1994.

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

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1236, 2003.

H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96 (456), 2001.

L. F. James. Large sample asymptotics for the two-parameter Poisson–Dirichlet process. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pages 187–199. Institute of Mathematical Statistics, 2008.

L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *The Annals of Statistics*, pages 1281–1296, 2001.

G. H. Jang, J. Lee, and S. Lee. Posterior consistency of species sampling priors. *Statistica Sinica*, 20(2):581, 2010.

Y. Ji, T. Lin, and H. Zha. CDP mixture models for data clustering. In *20th International Conference on Pattern Recognition (ICPR)*, pages 637–640. IEEE, 2010.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Ser. A*, 62(1): 49–66, 2000.

S. Khazaei, J. Rousseau, and F. Balabdaoui. Nonparametric Bayesian estimation of densities under monotonicity constraint. *(Preprint)*, 2012.

A. W. Knapp. *Basic Real Analysis*. Birkhäuser, 2005.

S. G. Krantz. *Function Theory of Several Complex Variables*. AMS Chelsea Publishing, Providence, 1992.

W. Kruijer. *Convergence Rates in Nonparametric Bayesian Density Estimation*. PhD thesis, Department of Mathematics, Vrije Universiteit Amsterdam, 2008.

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

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

A. D. Leaché and M. K. Fujita. Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). *Proceedings of the Royal Society B: Biological Sciences*, 277 (1697):3071–3077, 2010.

B. G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20 (3):1350–1360, 1992.

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

A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

A. Lijoi, I. Prünster, and S. G. Walker. Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547, 2008.

J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427): 958–966, 1994.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

E. D. Lorenzen, P. Arctander, and H. R. Siegismund. Regional genetic structuring and evolutionary history of the impala Aepyceros melampus. *Journal of Heredity*, 97(2): 119–132, 2006.

S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

S. N. MacEachern. Computational methods for mixture of Dirichlet process models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 23–43. Springer, 1998.

S. N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, 1999.

S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

J. W. Miller and M. T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems, Vol. 26*, 2013.

R. M. Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

X. L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

A. Nobile. *Bayesian Analysis of Finite Mixture Distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1994.

A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

A. Onogi, M. Nurimoto, and M. Morita. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12(1):263, 2011.

E. Otranto and G. M. Gallo. A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.

D. Pati, D. B. Dunson, and S. T. Tokdar. Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472, 2013.

J. Pella and M. Masuda. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(3):576–596, 2006.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

D. B. Phillips and A. F. M. Smith. Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, pages 215–239. Springer, 1996.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

J. Pitman. *Combinatorial Stochastic Processes*. Springer–Verlag, Berlin, 2006.

J. Pitman and M. Yor. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

C. M. Richards, G. M. Volk, A. A. Reilley, A. D. Henk, D. R. Lockwood, P. A. Reeves, and P. L. Forsline. Genetic diversity and population structure in Malus sieversii, a wild progenitor species of domesticated apple. *Tree Genetics & Genomes*, 5(2):339–347, 2009.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–792, 1997.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

T. Sapatinas. Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, 47(3):447–459, 1995.

C. Scricciolo. Adaptive Bayesian density estimation using Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *arXiv:1210.8094*, 2012.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *The Annals of Statistics*, 28 (1):40–74, 2000.

Y. Tang and S. Ghosal. Posterior consistency of Dirichlet mixtures for estimating a transition density. *Journal of Statistical Planning and Inference*, 137(6):1711–1726, June 2007.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963.

S. T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

S. G. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746, 2007.

M. West. Hyperparameter estimation in Dirichlet process mixture models. *ISDS Discussion Paper #92-A03, Duke University*, 1992.

M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman and A. F. Smith, editors, *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pages 363–386. Wiley, 1994.

M.-J. Woo and T. N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476), 2006.

M.-J. Woo and T. N. Sriram. Robust estimation of mixture complexity for count data. *Computational Statistics and Data Analysis*, 51(9):4379–4392, 2007.

Y. Wu and S. Ghosal. The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419, November 2010.

E. P. Xing, K. A. Sohn, M. I. Jordan, and Y. W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1049–1056, 2006.