INCONSISTENCY OF PITMAN–YOR PROCESS MIXTURES FOR THE NUMBER OF COMPONENTS

By Jeffrey W. Miller[*] and Matthew T. Harrison[*]

Brown University

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).

1. Introduction.

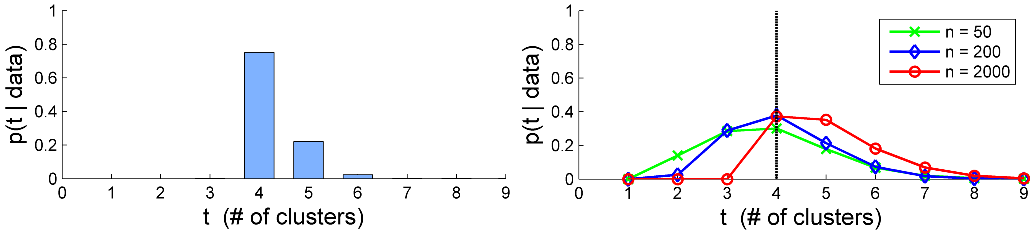
1.1. *A motivating example.* In population genetics, determining the "population structure" is an important step in the analysis of sampled data. As an illustrative example, consider the impala, a species of antelope in southern Africa. Impalas are divided into two subspecies: the common impala occupying 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, Arctander and Siegismund (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, Stephens and Donnelly, 2000) which, in the simplest version, models the data as a finite mixture, with each component in the mixture corresponding to a dis-

[*]Supported in part by NSF grant DMS-1007593 and DARPA contract FA8650-11-1-715.

AMS 2000 subject classifications: Primary 62G20; secondary 62G05.

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



(a) Posterior for impala data

(b) Posterior for Gaussian data

Fig 1: Estimated DPM posterior distribution of the number of clusters: (a) For the impala data of Lorenzen et al. ($n = 216$ datapoints). Our empirical results, shown here, agree with those of Huelsenbeck and Andolfatto. (b) For bivariate Gaussian data from a four-component mixture; see Figure 2. Each plot is the average over 10 independently-drawn datasets. (Lines drawn for visualization purposes only.) (For (a) and (b), estimates were made via Gibbs sampling, with 10^4 burn-in sweeps and 10^5 sample sweeps.)

tinct 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, as is common practice, 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 datapoints 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 prop-

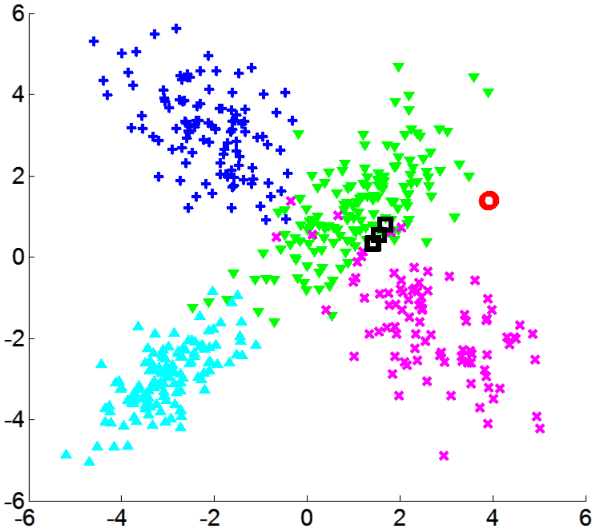


Fig 2: A typical partition sampled from the posterior of a Dirichlet process mixture of bivariate Gaussians, on simulated data from a four-component mixture. Different clusters have different marker shapes (+, ×, ▽, △,○, □) and different colors. Note the tiny "extra" clusters (○ and □), in addition to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

erties; see Section 1.3. However, in several applications, investigators have drawn inferences from the posterior on the number 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), 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 (and in particular, there are infinitely many components), 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, Figures 1(b) and 2 display results for data from a mixture of two-dimensional Gaussians with four components. Partitions sampled from the posterior often have tiny "extra" clusters, and the posterior on the number of clusters does not appear to be concentrating as the number of datapoints n increases. 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?

1.2. *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 1.1. *Consider a Pitman–Yor process mixture with component distributions from one of the following families:*

(a) $\mathrm{Normal}(\mu, \Sigma)$ *(multivariate Gaussian),*
(b) $\mathrm{Exponential}(\theta)$*,*
(c) $\mathrm{Gamma}(a, b)$*,*
(d) $\mathrm{Log\text{-}Normal}(\mu, \sigma^2)$*, or*
(e) $\mathrm{Weibull}(a, b)$ *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,*

$$\limsup_{n\to\infty} p(T_n = t \mid X_1, \dots, X_n) < 1$$

with probability 1*.*

This is implied by Theorems 3.4, 4.1, and 6.2. 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 6.2 covers a large class of exponential families, and Theorem 4.1 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.

Regarding the exception of $t = N$ when $\sigma < 0$ in Corollary 1.1: 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 1.3 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.

1.3. *Discussion / related work.* We would like to emphasize that this inconsistency should not be viewed as a deficiency of Dirichlet process mixtures, 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 a logarithmic factor (see Scricciolo (2012), Ghosal (2010) and references therein). Further, Nguyen (2013) has recently shown that the posterior on the mixing distribution converges in the Wasserstein metric to the true mixing distribution. However, these 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 (2013a), a brief exposition 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 (see Jang, Lee and Lee (2010); Lijoi, Mena and Prünster (2007) and references therein), but this is very different from our situation — in a species sampling model, the observed data is drawn directly from a 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, Müller and Escobar (1994); Lartillot and Philippe (2004); Onogi, Nurimoto and Morita (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. Among researchers using DPMs for clustering, this is an annoyance that is often

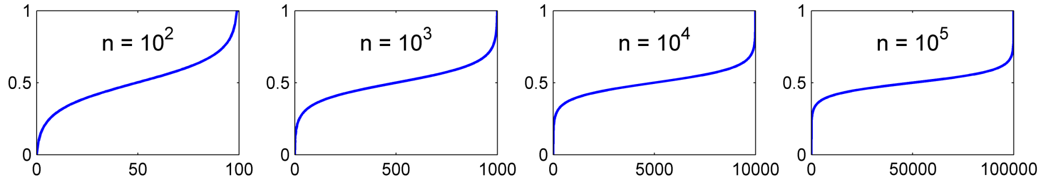


Fig 3: The cumulative distribution function of the conditional distribution of a_1 given that $t = 2$, for a Dirichlet process with $\vartheta = 1$. As n increases, the distribution becomes concentrated at the extremes.

dealt with by pruning such clusters — that is, by simply ignoring them when calculating statistics such as the number of clusters. 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.

However, if one is truly interested in estimating the number of components in a finite mixture, there is no need to resort to such measures — one can obtain posterior consistency by simply putting a prior on the number of components (Nobile, 1994). (It turns out that putting a prior on N in a PYM with $\sigma < 0$, $\vartheta = N|\sigma|$ is a special case of this (Gnedin and Pitman, 2006).) That said, it seems likely that such estimates will be severely affected by misspecification of the model, which is inevitable in most applications. Robustness to model misspecification seems essential for reliable estimation of the number of components, for real-world data.

1.4. *Intuition for the result.* To illustrate the intuition behind this inconsistency, consider a Dirichlet process with concentration parameter $\vartheta = 1$. (Similar reasoning applies for any Pitman–Yor process with $\sigma \geq 0$, but the $\sigma < 0$ case is somewhat different.) It is tempting to think that the prior on the number of clusters is the culprit, since (as is well-known) it diverges as $n \to \infty$. Surprisingly, this does not seem to be the main reason why inconsistency occurs.

Instead, the right intuition comes from examining the prior on partitions, *given* the number of clusters. The prior on ordered partitions $A = (A_1, \ldots, A_t)$ is $p(A) = (n!\,t!)^{-1} \prod_{i=1}^{t} (a_i - 1)!$, where t is the number of parts (i.e. clusters) and $a_i = |A_i|$ is the size of the ith part. (The $t!$ comes from uniformly permuting the parts; see Section 2.1.) Since there are $n!/(a_1! \cdots a_t!)$ such partitions with part sizes $(a_1, \ldots, a_t)$, the conditional distribution of the sizes $(a_1, \ldots, a_t)$ given t is proportional to $a_1^{-1} \cdots a_t^{-1}$ (subject to the constraint that $\sum a_i = n$). See Figure 3 for the case of $t = 2$. The key observation is that, for large n, this conditional distribution is heavily concentrated in the "corners", where one or more of the a_i's is small.

Pursuing this line of thought leads to the following startling fact: the probability of drawing a partition with $t + 1$ parts *and one or more of the a_i's equal to* 1 is, at least, the same order of magnitude (with respect to n) as the probability of drawing a partition with t parts. This leads to the basic idea of the proof — if the likelihood of the data is also the same order of magnitude, then the posterior probability of $t+1$ will not be too much smaller than that of t. Roughly speaking, the posterior will always find it reasonably attractive to split off one element as a singleton.

1.5. *Organization of the paper.* In Section 2, we define Gibbs partition mixture models, which includes Pitman–Yor and Dirichlet process mixtures as special cases. In Section 3, we prove a general inconsistency theorem for Gibbs partition 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 remainder of the paper proves the key lemma used in this application. In Section 7, we obtain certain inequalities involving the marginal density under an exponential family with conjugate prior. In Section 8, we prove the key lemma of Section 6: an inequality involving the marginal density of any sufficiently large subset of the data.

2. Model distribution.

A primary reason why inconsistency is possible in this situation is that the model is misspecified — that is, the data comes from a distribution that is not in the model class. Thus, 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.

Dirichlet process mixtures were introduced by Ferguson (1983) and Lo (1984) for the purpose of Bayesian density estimation, and were later made practical through the efforts of a number of authors (see Escobar and West (1998) and references therein). Pitman–Yor process mixtures (Ishwaran and James, 2003) are a generalization of DPMs based on the Pitman–Yor process (Pitman and Yor, 1997). The model we consider is, in turn, a generalization of PYMs based on the family of Gibbs partitions (Pitman, 2006).

2.1. *Gibbs partitions.* 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 $(A_1, \dots, A_t)$ 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.) on $\bigcup_{t=1}^{n} \mathcal{A}_t(n)$ of the form

$$p(A) = v_n(t) \prod_{i=1}^{t} w_n(|A_i|) \tag{2.1}$$

for $A \in \mathcal{A}_t(n)$, where $v_n : \{1, \dots, n\} \to [0, \infty)$ and $w_n : \{1, \dots, n\} \to [0, \infty)$. 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. Under this correspondence, any p.m.f. as in Equation 2.1 corresponds to a member of the class of "exchangeable partition probability functions", or EPPFs (Pitman, 2006). In particular, for any given n it yields an EPPF in "Gibbs form", and a random partition from such an EPPF is called a *Gibbs partition* (Pitman, 2006). (Note: We do not assume that, as n varies, the sequence of p.m.f.s in Equation 2.1 necessarily satisfies the marginalization property referred to as "consistency in distribution".)

For example, to obtain the partition distribution for a Dirichlet process (known as a Chinese restaurant process), we can choose

$$v_n(t) = \frac{\vartheta^t}{\vartheta_{n\uparrow 1}\ t!} \qquad \text{and} \qquad w_n(a) = (a-1)! \tag{2.2}$$

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, to obtain the partition distribution for a Pitman–Yor process, we can choose

$$v_n(t) = \frac{(\vartheta+\sigma)_{t-1\uparrow\sigma}}{(\vartheta+1)_{n-1\uparrow 1}\ t!} \qquad \text{and} \qquad w_n(a) = (1-\sigma)_{a-1\uparrow 1} \tag{2.3}$$

where either $\sigma \in [0, 1)$ and $\vartheta \in (-\sigma, \infty)$, or $\sigma \in (-\infty, 0)$ and $\vartheta = N|\sigma|$ for some $N \in \{1, 2, \dots\}$ (Ishwaran and James, 2003). When $\sigma = 0$, this reduces to the partition distribution of a 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 $\{1, \dots, N\}$.

2.2. *Gibbs partition mixtures.* Consider the hierarchical model

$$
\begin{aligned}
p(A,t) = p(A) &= v_n(t)\prod_{i=1}^t w_n(|A_i|), \\
p(\theta_{1:t}\mid A,t) &= \prod_{i=1}^t \pi(\theta_i), \\
p(x_{1:n}\mid \theta_{1:t},A,t) &= \prod_{i=1}^t \prod_{j\in A_i} p_{\theta_i}(x_j),
\end{aligned}
\tag{2.4}
$$

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 parametrized 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. (Of course, the distribution of x under $p_\theta(x)$ may be discrete, continuous, or neither, depending on the nature of λ.)

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

$$
m(x_J) = \int_\Theta \Big(\prod_{j\in J} p_\theta(x_j)\Big)\,\pi(\theta)\,d\theta, \tag{2.5}
$$

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 2.1. We refer to such a hierarchical model as a *Gibbs partition mixture model.*

(Note: This is, perhaps, a slight abuse of the term "Gibbs partition", since we allow v_n and w_n to vary arbitrarily with n.) In particular, it is a *Dirichlet process mixture model* when v_n and w_n are as in Equation 2.2, or more generally, a *Pitman–Yor process mixture model* when v_n and w_n are as in Equation 2.3.

We distinguish between the terms "component" and "cluster": a *component* of a mixture is one of the distributions used in it (e.g. p_{θ_i}), while a *cluster* is the set of indices of datapoints coming from a given component (e.g. A_i). The prior on the number of clusters under such a model is $p_n(t)=\sum_{A\in\mathcal{A}_t(n)}p(A)$. 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

$$
\begin{aligned}
p(x_{1:n}, T_n = t) &= \sum_{A \in \mathcal{A}_t(n)} \int p(x_{1:n}, \theta_{1:t}, A, t)\, d\theta_{1:t} \\
&= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^{t} \int \Big(\prod_{j \in A_i} p_{\theta_i}(x_j) \Big) \pi(\theta_i)\, d\theta_i \\
&= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^{t} m(x_{A_i}). \qquad (2.6)
\end{aligned}
$$

As usual, the posterior $p(T_n = t \mid x_{1:n})$ is not uniquely defined, since it can be modified arbitrarily on any subset of $\mathcal{X}^n$ having probability zero under the model distribution. For definiteness, we will employ the usual version of this posterior,

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

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

3. Inconsistency theorem.

The essential ingredients in the main theorem are Conditions 3.1 and 3.2 below. For each $n \in \{1, 2, \dots\}$, consider a partition distribution as in Equation 2.1. For $n > t \geq 1$, define

$$
c_{w_n} = \max_{a \in \{2,\dots,n\}} \frac{w_n(a)}{a\, w_n(a-1)\, w_n(1)} \quad \text{and} \quad c_{v_n}(t) = \frac{v_n(t)}{v_n(t+1)},
$$

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

Condition 3.1. *Assume* $\limsup_{n\to\infty} c_{w_n} < \infty$ *and* $\limsup_{n\to\infty} c_{v_n}(t) < \infty$, *given some particular* $t \in \{1, 2, \dots\}$.

For Pitman–Yor processes, Condition 3.1 holds for all relevant values of t, by Proposition 3.3 below. Now, consider a collection of single-cluster marginals $m(\cdot)$ as in Equation 2.5. Given $n \geq t \geq 1$, $x_1, \dots, x_n \in \mathcal{X}$, 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 $j \in \{1, \dots, n\}$ 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 3.2. *Given a sequence of random variables* $X_1, X_2, \dots \in \mathcal{X}$, *a collection of single-cluster marginals* $m(\cdot)$, *and* $t \in \{1, 2, \dots\}$, *assume*

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

Note that Condition 3.1 involves only the partition distributions, while Condition 3.2 involves only the data distribution and the single-cluster marginals.

PROPOSITION 3.3. *Consider a Pitman–Yor process. If* $\sigma \in [0,1)$ *and* $\vartheta \in (-\sigma,\infty)$ *then Condition 3.1 holds for any* $t \in \{1,2,\dots\}$*. If* $\sigma \in (-\infty,0)$ *and* $\vartheta = N|\sigma|$*, then it holds for any* $t \in \{1,2,\dots\}$ *except* N*.*

PROOF. This is a simple calculation. See Appendix A. □

THEOREM 3.4. *Let* $X_1, X_2, \dots \in \mathcal{X}$ *be a sequence of random variables (not necessarily i.i.d.). Consider a Gibbs partition mixture model. For any* $t \in \{1,2,\dots\}$*, if Conditions 3.1 and 3.2 hold, then*

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

If, further, the sequence $X_1, X_2, \dots$ *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. This follows easily from Lemma 3.5 below. See Appendix A. □

LEMMA 3.5. *Consider a Gibbs partition mixture model. Let* $n > t \geq 1$*,* $x_1,\dots,x_n \in \mathcal{X}$*, and* $c \in [0,\infty)$*. If* $\varphi_t(x_{1:n},c) > t/n$*,* $c_{w_n} < \infty$*, and* $c_{v_n}(t) < \infty$*, 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 $C_t(x_{1:n},c) = t\,c\,c_{w_n} c_{v_n}(t)/(\varphi_t(x_{1:n},c) - t/n)$*.*

PROOF. To simplify notation, let us denote $\varphi = \varphi_t(x_{1:n},c)$, $C = C_t(x_{1:n},c)$, and $S_A = S_A(x_{1:n},c)$ for $A \in \mathcal{A}_t(n)$. Given $J \subset \{1,\dots,n\}$ such that $|J| \geq 1$, define

$$h_J = w_n(|J|)\, m(x_J).$$

For $A \in \mathcal{A}_t(n)$, let $R_A = S_A \smallsetminus \big(\bigcup_{i:|A_i|=1} A_i\big)$, that is, R_A consists of those $j \in S_A$ such that the size of the part A_ℓ containing j is greater than 1. Note that

$$|R_A| \geq |S_A| - t \geq n\varphi - t > 0. \tag{3.1}$$

For any $j \in R_A$, the part A_ℓ containing j satisfies

$$\begin{aligned} h_{A_\ell} &= w_n(|A_\ell|)\, m(x_{A_\ell}) \\ &\leq c_{w_n}\, |A_\ell|\, w_n(|A_\ell| - 1)\, w_n(1)\, c\, m(x_{A_\ell \smallsetminus j})\, m(x_j) \\ &\leq n\, c\, c_{w_n}\, h_{A_\ell \smallsetminus j}\, h_j. \end{aligned} \tag{3.2}$$

Given $j \in R_A$, define $B(A,j)$ to be the element B of $\mathcal{A}_{t+1}(n)$ such that $B_i = A_i \smallsetminus j$ for $i = 1, \dots, t$, and $B_{t+1} = \{j\}$ (that is, remove j from whatever part it belongs to, and make $\{j\}$ the $(t+1)^{\text{th}}$ part). Define

$$\mathcal{Y}_A = \big\{B(A,j) : j \in R_A\big\}.$$

Now, using Equations 3.1 and 3.2, for any $A \in \mathcal{A}_t(n)$ we have

$$\begin{aligned}\prod_{i=1}^{t} h_{A_i} &= \frac{1}{|R_A|}\sum_{\ell=1}^{t}\sum_{j\in R_A\cap A_\ell} h_{A_\ell}\prod_{i\neq\ell} h_{A_i} \\ &\leq \frac{1}{n\varphi - t}\sum_{\ell=1}^{t}\sum_{j\in R_A\cap A_\ell} n\,c\,c_{w_n}\,h_{A_\ell\smallsetminus j}\,h_j\prod_{i\neq\ell} h_{A_i} \\ &= \frac{c\,c_{w_n}}{\varphi - t/n}\sum_{j\in R_A}\prod_{i=1}^{t+1} h_{B_i(A,j)} \\ &= \frac{c\,c_{w_n}}{\varphi - t/n}\sum_{B\in\mathcal{A}_{t+1}(n)}\Big[\prod_{i=1}^{t+1} h_{B_i}\Big] I(B\in\mathcal{Y}_A).\end{aligned} \tag{3.3}$$

For any $B \in \mathcal{A}_{t+1}(n)$,

$$\#\big\{A \in \mathcal{A}_t(n) : B \in \mathcal{Y}_A\big\} \leq t, \tag{3.4}$$

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

$$\begin{aligned}p(x_{1:n}, T_n = t) &\overset{\text{(a)}}{=} \sum_{A\in\mathcal{A}_t(n)} p(A)\prod_{i=1}^{t} m(x_{A_i}) \\ &\overset{\text{(b)}}{=} \sum_{A\in\mathcal{A}_t(n)} v_n(t)\prod_{i=1}^{t} h_{A_i} \\ &\overset{\text{(c)}}{\leq} \frac{c\,c_{w_n}}{\varphi - t/n}\,v_n(t)\sum_{A\in\mathcal{A}_t(n)}\sum_{B\in\mathcal{A}_{t+1}(n)}\Big[\prod_{i=1}^{t+1} h_{B_i}\Big] I(B\in\mathcal{Y}_A) \\ &= \frac{c\,c_{w_n}}{\varphi - t/n}\,v_n(t)\sum_{B\in\mathcal{A}_{t+1}(n)}\Big[\prod_{i=1}^{t+1} h_{B_i}\Big]\#\big\{A\in\mathcal{A}_t(n) : B\in\mathcal{Y}_A\big\} \\ &\overset{\text{(d)}}{\leq} \frac{c\,c_{w_n}c_{v_n}(t)}{\varphi - t/n}\,v_n(t+1)\sum_{B\in\mathcal{A}_{t+1}(n)}\Big[\prod_{i=1}^{t+1} h_{B_i}\Big] t \\ &= \frac{t\,c\,c_{w_n}c_{v_n}(t)}{\varphi - t/n}\sum_{B\in\mathcal{A}_{t+1}(n)} p(B)\prod_{i=1}^{t+1} m(x_{B_i}) \\ &= C\ p(x_{1:n}, T_n = t+1),\end{aligned}$$

where (a) is from Equation 2.6, (b) is from Equation 2.4 and the definition of h_J above, (c) follows from Equation 3.3, and (d) follows from Equation 3.4.

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

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

□

4. Application to discrete or bounded cases. By Theorem 3.4, 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 4.1. *Consider a family of densities $\{p_\theta : \theta \in \Theta\}$ on $\mathcal{X}$ along with a prior π on Θ and the resulting collection of single-cluster marginals $m(\cdot)$ as in Equation 2.5. 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) $\liminf_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} I(X_j \in U) > 0$ *with probability* 1, *and*

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

then Condition 3.2 holds for all $t \in \{1, 2, \ldots\}$.

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

$$m(x_J) = \int_\Theta p_\theta(x_j)\Big[\prod_{i\in J\smallsetminus j} p_\theta(x_i)\Big]\pi(\theta)\,d\theta \le c\, m(x_j) m(x_{J\smallsetminus j}). \tag{4.1}$$

Thus, letting $R(x_{1:n}) = \big\{j \in \{1, \ldots, n\} : x_j \in U\big\}$, 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,

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

□

The preceding theorem covers a fairly wide range of cases; here are some examples. Consider a model with $\{p_\theta\}$, π, λ, and $m(\cdot)$, as in Section 2.

(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 (1) and (2) of Theorem 4.1 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 1.1.

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 (1) and (2) are satisfied with $U = \{x_0\}$. (Note that $\lambda(x_0) > 0$ implies $p_\theta(x_0) \leq 1/\lambda(x_0)$ 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 (1) is satisfied due to item (a), and condition (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 4.1: condition (2) is not satisfied in some important cases, such as multivariate Gaussians with unrestricted covariance. Showing that Condition 3.2 still holds, for many exponential families at least, is the objective of the remainder of the paper.

5. Exponential families and conjugate priors.

5.1. *Exponential families.* In this section, we make the usual definitions for exponential families and state the regularity conditions to be assumed. 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 reparametrizing if necessary, and choosing λ appropriately. We will assume the following (very mild) regularity conditions.

CONDITIONS 5.1. *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 Conditions 5.1, including multivariate Gaussian, Gamma, Poisson, Exponential, 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 B 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

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

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

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

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

Definition 5.2. We will say that an exponential family with conjugate prior is *well-behaved* if it takes the form above, satisfies Conditions 5.1, and has $(\xi,\nu) \in \Xi$.

6. Application to exponential families. In this section, we apply Theorem 3.4 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.

Conditions 6.1. *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 $L = \{x \in \mathbb{R}^k : x^{\mathtt{T}}y = b\}$ for some $y \in \mathbb{R}^k \smallsetminus \{0\}$, $b \in \mathbb{R}$. In Theorem 6.2 below, it is assumed that the data comes from a distribution P satisfying Conditions 6.1. In Proposition 6.3, we give some simple conditions under which, if P is a finite mixture from the exponential family under consideration, then Conditions 6.1 hold.

The following theorem follows almost immediately from Lemma 8.4, the proof of which will occupy most of the remainder of the paper.

Theorem 6.2. *Consider a well-behaved exponential family with conjugate prior (as in Definition 5.2), along with the resulting collection of single-cluster marginals $m(\cdot)$. Let P be a probability measure on $\mathcal{X}$ satisfying Conditions 6.1 (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 3.2 holds for any $t \in \{1,2,\ldots\}$.*

PROOF. Let $t \in \{1, 2, \dots\}$ and choose c according to Lemma 8.4 with $\beta = 1/t$. We will show that for any $n > t$, if the event of Lemma 8.4 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 $x_1, \dots, x_n \in \mathcal{X}$, and assume the event of Lemma 8.4 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 $R_A \subset A_\ell$ such that $|R_A| \geq \frac{1}{2}|A_\ell|$ and for any $j \in R_A$, $m(x_{A_\ell}) \leq c\, m(x_{A_\ell \smallsetminus j})\, m(x_j)$. Thus, $R_A \subset S_A(x_{1:n}, c)$, hence $|S_A(x_{1:n}, c)| \geq |R_A| \geq \frac{1}{2}|A_\ell| \geq n/(2t)$. Since $A \in \mathcal{A}_t(n)$ was arbitrary, $\varphi_t(x_{1:n}, c) \geq 1/(2t)$. □

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 6.3 below, that if P is a finite mixture from the same family then P satisfies Conditions 6.1:

(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 4.1, these results are fairly comprehensive.

PROPOSITION 6.3. *Consider an exponential family $\{P_\theta : \theta \in \Theta\}$ satisfying Conditions 5.1. If $X \sim P = \sum_{i=1}^t \pi_i P_{\theta(i)}$ for some $\theta(1), \dots, \theta(t) \in \Theta$ and some $\pi_1, \dots, \pi_t \geq 0$ 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 exponential family is continuous (that is, 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 $s_1, \dots, s_k$ is real analytic), then

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

PROOF. This is relatively straightforward; see the Supplementary Material. □

Sometimes, Condition 6.1(3) will be satisfied even when Proposition 6.3 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 B.1(8) in the Appendix).

7. Marginal inequalities. Consider a well-behaved exponential family with conjugate prior (as in Definition 5.2). In this section, we use some simple bounds on the Laplace approximation (see Appendix C) to prove certain inequalities involving the marginal density (from Equation 2.5),

$$m(x_{1:n}) = \int_\Theta \Big(\prod_{j=1}^n p_\theta(x_j)\Big)\pi_{\xi,\nu}(\theta)\,d\theta$$

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. In contrast, our application of it is considerably more subtle. For our purposes, it is necessary to show that the approximation is good not only in the i.i.d. case, but in fact whenever the sufficient statistics are not too extreme.

We make extensive use of the exponential family properties in Appendix B, 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

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

and note that $\theta_\mu = \kappa'^{-1}(\mu)$ (Proposition B.1). $\mathcal{L}$ 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 $\mathcal{L}(\mu) = \theta_\mu^{\mathsf{T}}\mu - \kappa(\theta_\mu)$, $\theta_\mu = \kappa'^{-1}(\mu)$, and both κ and κ'^{-1} are C^∞ smooth). Define

$$\mu_{x_{1:n}} = \frac{\xi + \sum_{j=1}^n s(x_j)}{\nu + n} \tag{7.1}$$

(cf. Equation B.1), 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 $s : \mathcal{X} \to \mathbb{R}^k$. The first of the two results of this section 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 7.1. *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

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

(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 B.2) are continuous on $\mathcal{M}$. Define

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

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 C.2 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 B.1, 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,

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

Thus, combining the preceding three displayed equations,

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

The second result of this section is an inequality involving a product of marginals.

Proposition 7.2 (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 $b = \nu + |B|$.

Proof. 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 7.1 to U', and

$$C_0 = \nu \sup_{y\in U'} |(\xi/\nu - y)^{\mathtt{T}} \mathcal{L}'(y)| + \nu \sup_{y \in U'} |\mathcal{L}(y)| < \infty. \tag{7.2}$$

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 $\rho \in (0,1)$,

$$\mathcal{L}(y) \leq \mathcal{L}((1-\rho)y + \rho z) - \rho(z - y)^{\mathtt{T}} \mathcal{L}'(y). \tag{7.3}$$

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

$$(1-\rho)y + \rho z = \frac{2\xi + \sum_{j=1}^n s(x_j)}{2\nu + n} = \frac{a\mu_{x_A} + b\mu_{x_B}}{a + b}. \tag{7.4}$$

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,

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

$$\overset{\text{(b)}}{\leq} (a+b)\mathcal{L}\Big(\frac{a\mu_{x_A} + b\mu_{x_B}}{a+b}\Big) + C_0$$
$$\overset{\text{(c)}}{\leq} a\mathcal{L}(\mu_{x_A}) + b\mathcal{L}(\mu_{x_B}) + C_0,$$

where (a) is by Equation 7.3, (b) is by Equations 7.2 and 7.4, 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,

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

8. Marginal inequality for subsets of the data.

In this section, we prove Lemma 8.4, the key lemma used in the proof of Theorem 6.2. First, we need a few supporting results.

Given $y_1, \ldots, y_n \in \mathbb{R}^\ell$ (for some $\ell > 0$), $\beta \in (0,1]$, and $U \subset \mathbb{R}^\ell$, define

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

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

Lemma 8.1 (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 $Y_1, Y_2, \ldots \overset{\text{iid}}{\sim} Q$, 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. For details, see the Supplementary Material. □

Proposition 8.2. *Let $Z_1, Z_2, \ldots \in \mathbb{R}^k$ 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 $A \subset \{1,\ldots,n\}$ 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, $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$. □

In the following, $\mu_x = (\xi + s(x))/(\nu+1)$, as in Equation 7.1.

PROPOSITION 8.3. *Consider a well-behaved exponential family with conjugate prior. Let P be a probability measure on $\mathcal{X}$ such that $\mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$ when $X \sim P$. Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. 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 $y \in \bar{\mathcal{M}}$, $z \in \mathcal{M}$, 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 $U_0 = \{(\xi+y)/(\nu+1) : y \in \bar{\mathcal{M}}\}$ is contained in $\mathcal{M}$. Note that U_0 is closed and $\mathbb{P}(\mu_X \in U_0) = \mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$. 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 8.2 with $Z_j = \mu_{X_j}$, we have $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$. □

LEMMA 8.4. *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 Conditions 6.1 (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 Conditions 6.1 hold by assumption, then by Lemma 8.1 (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 8.3 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 7.2 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 $|J| \geq \beta n$. □

APPENDIX A

Proof of Proposition 3.3. There are two cases: (A) $\sigma \in [0,1)$ and $\vartheta > -\sigma$, or (B) $\sigma < 0$ and $\vartheta = N|\sigma|$. In either case, $\sigma < 1$, so

$$\frac{w_n(a)}{a w_n(a-1) w_n(1)} = \frac{1 - \sigma + a - 2}{a} \leq \frac{1-\sigma}{2} + 1$$

whenever $n \geq 2$ and $a \in \{2, \ldots, n\}$, and hence $\limsup_n c_{w_n} < \infty$.

For any $n > t \geq 1$, in case (A) we have

$$\frac{v_n(t)}{v_n(t+1)} = \frac{t+1}{\vartheta + t\sigma},$$

and the same holds in case (B) if also $t < N$. Meanwhile, whenever $N < t < n$ in case (B), $v_n(t)/v_n(t+1) = 0/0 = 0$ by convention. Therefore, $\limsup_n c_{v_n}(t) < \infty$ in either case, for any $t \in \{1, 2, \ldots\}$ except $t = N$ in case (B). □

Proof of Theorem 3.4. Let $t \in \{1, 2, \ldots\}$, and assume Conditions 3.1 and 3.2 hold. Let $x_1, x_2, \ldots \in \mathcal{X}$, 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 $\limsup_n c_{w_n} < \alpha$ and $\limsup_n c_{v_n}(t) < \alpha$. 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 $c_{w_n} < \alpha$, $c_{v_n}(t) < \alpha$, and $\varphi_t(x_{1:n}, c) > \varepsilon$. Then by Lemma 3.5, for any $n > N$,

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

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), then $\limsup_n p(T_n = t \mid x_{1:n}) < 1$. □

APPENDIX B: EXPONENTIAL FAMILY PROPERTIES

We note some well-known properties of exponential families satisfying Conditions 5.1. 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 B.1. *If Conditions 5.1 are 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 $\bar{\mathcal{M}} = C_\lambda(s)$, 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 $x_1, \ldots, x_n \in \mathcal{X}$ and $\theta \in \mathbb{R}^k$ is

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

where $\mu_{x_{1:n}} = (\xi + \sum_{j=1}^n s(x_j))/(\nu+n)$. The marginal density, defined as in Equation 2.5, is

$$m(x_{1:n}) = \exp\Big(\psi\big(\xi + \textstyle\sum s(x_j),\, \nu+n\big) - \psi(\xi,\nu)\Big) \tag{B.2}$$

when this quantity is well-defined.

Proposition B.2. *If Conditions 5.1 are satisfied, then:*

1. *$\psi(\xi,\nu)$ is finite and C^∞ smooth on Ξ,*
2. *if $s(x_1),\ldots,s(x_n)\in S_\lambda(s)$ 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 $x_1,\ldots,x_n\in\mathcal{X}$.*

Proof. (1) For finiteness, see Diaconis and Ylvisaker (1979), Theorem 1. Smoothness holds for the same reason that κ is smooth (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 (B.1(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 $X_1,\ldots,X_n\overset{\text{iid}}{\sim}P_\theta$ then $s(X_1),\ldots,s(X_n)\in S_\lambda(s)$ almost surely, and thus $(\xi+\sum s(X_j),\,\nu+n)\in\Xi$ (a.s.) by (2). By Equations B.1 and B.2, 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 $x_1,\ldots,x_n\in\mathcal{X}$, we have $(\xi+\sum s(x_j),\,\nu+n)\in\Xi$ by (2). Thus, by (1) and Equation B.2, $m(x_{1:n})<\infty$. □

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.

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 C.1 is simply a multivariate version of the bounds given by De Bruijn (1970), and Corollary C.2 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

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

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

Lemma C.1. *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* $B_\varepsilon(x_0) \subset E$, $0 < \delta \leq \inf\{f(x) - f(x_0) : x \in E \smallsetminus B_\varepsilon(x_0)\}$, *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* $0 < s \leq t$ *we have*

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

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

Remark. 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. This is a straightforward application of Taylor's theorem; see the Supplementary Material. □

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 $\lambda_*(A)^k \leq |A| \leq \lambda^*(A)^k$.

Corollary C.2. *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 C.1 (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 $t \geq 1$. □

ACKNOWLEDGMENTS

We would like to thank Stu Geman for raising this question. We would also like to thank Erik Sudderth, Mike Hughes, Tamara Broderick, Annalisa Cerquetti, Steve MacEachern, Dan Klein, and Dahlia Nadkarni for helpful conversations.

SUPPLEMENTARY MATERIAL

Supplementary Material: Supplement to "Inconsistency of Pitman–Yor process mixtures for the number of components"
(doi: TBD; .pdf). We have placed some technical proofs in a supplementary document, Miller and Harrison (2013b).

REFERENCES

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

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

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

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

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

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

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

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

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

Hoffmann-Jørgensen, J. (1994). *Probability with a view toward statistics* **2**. Chapman & Hall.

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

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

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

Knapp, A. W. (2005). *Basic real analysis.* Birkhäuser.

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

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

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

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

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

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

Miller, J. W. and Harrison, M. T. (2013a). A simple example of Dirichlet process mixture inconsistency for the number of components arXiv:1301.2708 [math.ST].

Miller, J. W. and Harrison, M. T. (2013b). Supplement to "Inconsistency of Pitman–Yor process mixtures for the number of components".

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

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

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

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

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

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Springer-Verlag, Berlin.

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

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

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

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

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

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

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

Teicher, H. (1963). Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics* **34** 1265–1269.

West, M., Müller, P. and Escobar, M. D. (1994). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

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

Division of Applied Mathematics
Brown University
Providence, RI 02912
E-mail: Jeffrey_Miller@Brown.edu
Matthew_Harrison@Brown.edu