Applications and Computation of Stateful Polya Trees

by

Jonathan Christensen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Supervisor

_____
Scott Schmidler

_____
Surya Tokdar

_____
Cliburn Chan

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2017

Abstract

Applications and Computation of Stateful Polya Trees

by

Jonathan Christensen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Supervisor

_____
Scott Schmidler

_____
Surya Tokdar

_____
Cliburn Chan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2017

Abstract

Polya trees are a class of nonparametric priors on distributions which are able to model absolutely continuous distributions directly, rather than modeling a discrete distribution over parameters of a mixing kernel to obtain an absolutely continuous distribution. The Polya tree discretizes the state space with a recursive partition, generating a distribution by assigning mass to the child elements at each level of the recursive partition according to a Beta distribution. Stateful Polya trees are an extension of the Polya tree where each set in the recursive partition has one or more discrete state variables associated with it. We can learn the posterior distributions of these state variables along with the posterior of the distribution. State variables may be of interest in their own right, or may be nuisance parameters which we use to achieve more flexible models but wish to integrate out in the posterior. We discuss the development of stateful Polya trees and discuss the Hierarchical Adaptive Polya Tree, which uses state variables to flexibly model the concentration parameter of Polya trees in a hierarchical Bayesian model. We also consider difficulties with the use of marginal likelihoods to determine posterior probabilities of states.

For Kathryn

Contents

List of Figures

Acknowledgements

My wife, Kathryn, for all her patience and support.

My parents, for teaching me to think carefully and thoroughly.

My advisor, Li Ma, for sharing his ideas and encouraging me to share mine.

Mike West and Surya Tokdar, for thoughtful conversations and encouragement when I needed it most.

The entire Duke Statistical Science community, for the fantastic environment—both intellectual and personal—that I have enjoyed during my years here.

1

Introduction

The Polya tree is a nonparametric prior on distributions introduced by Freedman (1963) as a special case of the tail-free prior. Tail-free priors are a class of priors on distributions in which a recursive partition $\mathcal{P}$ is placed on the sample space, and mass assigned conditionally at each level of the recursive partition according to a sequence of random variables taking values on the interval $[0, 1]$. The Polya tree is given by the case where the the recursive partition is a binary partition tree and the corresponding random variables are given Beta distributions, although it can be generalized to partitions that are not binary, with corresponding Dirichlet random variables.

The Polya tree includes as a special case the Dirichlet process of Ferguson (1973), though its specification is considerably more general. While the Dirichlet process almost surely generates discrete distributions, with appropriate specification of the prior parameters the Polya tree almost surely generates absolutely continuous distributions (Ferguson, 1974). This is made possible by the more flexible nature of the Polya tree's concentration parameter, which is infinite dimensional, in contrast to the Dirichlet process's single-dimensional concentration parameter.

1.1 Reviewing the Polya tree construction

We begin with a brief sketch of the Polya tree; the reader interested in the mathematical details should refer to Mauldin et al. (1992); Lavine (1992, 1994). The Polya tree consists of an infinite recursive partition $\mathcal{A}$ of the sample space and a corresponding infinite sequence of Beta-distributed random variables which assign mass to the various regions $A \in \mathcal{A}$ of the partition. Figure 1.1 illustrates the partitioning sequentially. In this illustration our sample space is the interval $(0, 1]$ on the real line, and our prior mean, shown in the first pane, is the uniform distribution on that interval. At each level of the recursive partition we cut each region in half at the midpoint. Although arbitrary partitions may be used, the dyadic partition described here is convenient and is often used as a default partition in the absence of a compelling reason to use a different one. Other partitions may be more convenient in the presence, for example, of censored data (Muliere and Walker, 1997). The second pane shows the result after the first cut and mass allocation, in this case the majority of the mass having been allocated to the right-hand side. In the next step we cut each of the two regions from the second pane in half again, and assign the probability mass to each to its children according to a Beta-distributed random variable. This results in four regions, shown in the third pane, which are again cut and mass distributed in the fourth pane. The process continues indefinitely.

We denote the Polya tree prior as $Q \sim \mathrm{PT}(Q_0, \boldsymbol{\nu})$, where Q_0 is the centering distribution and $\boldsymbol{\nu}$ is an (infinite dimensional) concentration parameter describing how much Q is expected to vary from Q_0. The parameters of the sequence of Beta distributions from which the mass allocations are drawn are derived very simply from Q_0 and $\boldsymbol{\nu}$. For an arbitrary region A belonging to the recursive partition $\mathcal{A}$, the fraction of the mass allocated to the left child A_ℓ of A is given by the random

variable

$$\theta(A) \sim \text{Beta}(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A)),$$

where $\theta_0(A) = Q_0(A_\ell)/Q_0(A)$, with the remainder of the mass allocated to the right child A_r. Q_0 thus determines the mean of the mass allocations (and hence the expectation of the resulting density), while $\boldsymbol{\nu} = \{\nu(A) : A \in \mathcal{A}\}$ controls the variation of the mass allocations, and hence the dispersion of Q around Q_0. Alternatively, $\boldsymbol{\nu}$ controls the strength of the shrinkage of the posterior mean density from the empirical process towards Q_0. With an appropriate choice of $\boldsymbol{\nu}$, the Polya tree prior almost surely generates an absolutely continuous distribution (Kraft, 1964).

Polya trees and extensions have been widely used as nonparametric priors for absolutely continuous distributions. We give a few notable examples.

With appropriate choice of the recursive partition structure the Polya tree is a convenient nonparametric model for survival analysis (Muliere and Walker, 1997). The construction is as follows: suppose we have right-censored positive observations, and let $0 < a_1 < \cdots < a_n$ be a set containing all observed censoring times (and possibly other points). We construct an asymmetric recursive partition of the positive half-line as

$$\begin{aligned}
[0, \infty) &= [0, a_1) \cup [a_1, \infty) \\
[a_1, \infty) &= [a_1, a_2) \cup [a_2, \infty) \\
&\vdots \\
[a_{n-1}, \infty) &= [a_{n-1}, a_n) \cup [a_n, \infty).
\end{aligned}$$

Because the censoring times correspond to left endpoints of sets in the recursive partition with right-endpoint at infinity, the uncertainty due to censoring is easily integrated out from the posterior. Both censored and uncensored observations are counted in every region which we *know* them to belong to; censored observations do not effect the likelihood in any region to which their relationship is uncertain.

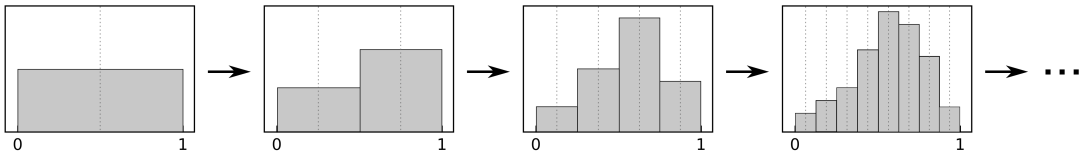

FIGURE 1.1: An illustration of the recursive partitioning and probability allocation of the standard Polya tree.

This construction depends on *a priori* knowledge of the censoring times. In some applications this may actually be the case, for example if participants will be dropped from a planned study at preset times, and no unplanned censoring occurs. More often, however, censoring times cannot be predicted; in this case, fitting the model of Muliere and Walker (1997) requires looking at the data to specify the prior, a practice which is philosophically questionable, if common. Neath (2003) resolves this concern by showing how to model censored data without specifying the partition structure to match censoring times, obtaining a posterior which is a mixture of Polya trees. He shows how to enumerate the mixture components and compute the probability associated with each, the remaining part of the problem being simply to compute the posterior distribution for each of the associated Polya trees, which is easily done. He illustrates the model with a small data set which is computationally tractable; unfortunately, as the number of censored observations and the resolution of the recursive partition grow, the number of mixture components becomes infeasible.

Zhao and Hanson (2011) expand the use of Polya trees in survival analysis by introducing a spatially dependent Polya tree. A Polya tree prior is placed on the baseline distribution in a proportional hazards model. Spatial dependence is induced on the mass assignment parameters of the tree; for the sake of computational tractability, the Beta distributions are approximated by Normal distributions. Markov-chain Monte Carlo is used to estimate posterior parameters, as the complexities of the model make the standard conjugate Polya tree posterior update impossible.

Paddock (2002) places a multidimensional Polya tree prior on the joint distribution of data which is observed with missing values. Rather than the canonical dyadic partition from the one-dimensional case, the multidimensional Polya tree described here cuts each region along every dimension, resulting in 2^d child regions. As in Neath (2003), the presence of missing data results in the posterior being a mixture of Polya trees rather than a single Polya tree; again the number of mixture com-

ponents quickly becomes intractable in all but the smallest applications. Paddock avoids enumerating the mixture components by describing a Markov-chain Monte Carlo sampler which draws from an approximation to the posterior, as even the true posterior is computationally difficult to draw from. Treating each missing value as a model parameter, a Gibbs sampler iterates between drawing values for the unobserved values and fitting and sampling from the resulting (conjugate) Polya tree, a Bayesian counterpart to the EM algorithm. The resulting draws from the joint posterior can be used to estimate the joint distribution directly or impute missing values.

This approach has a number of weaknesses, including the large number of degrees of freedom used by the multivariate Polya tree partitioning scheme, the difficulty in computation, and concerns over the accuracy of two levels of approximation—approximation of the posterior distribution to define the Gibbs sampler, and the Monte Carlo error of the sampler itself.

In addition to purely nonparametric models, Polya trees have been used to construct a number of semiparametric models. For example, Walker and Mallick (1997) use a Polya tree prior for the distribution of random effects in a generalized linear mixed model. A Gibbs sampler is used to alternate between drawing from the full conditional posterior for the random effects distribution (fitting and drawing from the conjugate Polya tree, as in Paddock (2002)) and drawing from the full conditional posterior distributions of the other model parameters. They illustrate the model with several applications, including one in which the distribution of random effects appears to be bimodal, a feature which the usual parametric analysis was incapable of capturing.

Similarly, Hanson and Johnson (2002) use a mixture of Polya trees to model the error distribution in a linear regression problem, under the constraint that the median of the error distribution is equal to zero. This constraint is easily enforced in

the canonical Polya tree partitioning system, unlike constraints on the mean of the distribution. Again a Gibbs sampler is used to iterate between sampling the error distribution from the conjugate Polya Tree full conditional posterior and sampling the other model parameters from their full conditional posteriors. The use of a nonparametric prior allows the model to capture an error distribution with considerable skewness, which would otherwise result in substantial bias in the estimation of other model parameters.

Other extensions have attempted to deal with the partition dependence which Polya trees (and all other tail-free priors, except the Dirichlet Process) suffer from. For example, Paddock et al. (2003) use a mixture of Polya trees with jittered partition points, while Nieto-Barajas and Müller (2012) introduce dependence in the mass assignments at a given depth of the tree.

Like all tail-free priors, the Polya tree prior is conjugate, and it allows easy computation of the marginal likelihood. This has made it popular as a nonparametric model for Bayesian hypothesis testing. Berger and Guglielmi (2001) use it as a nonparametric alternative to test a parametric null hypothesis against, though they find considerable sensitivity to the choice of prior parameters of the Polya tree. Holmes et al. (2015) use the Polya tree to test the null hypothesis that two samples come from the same distribution versus an alternative of two independent distributions, with each distribution having a Polya tree prior.

1.2 Stateful Polya Trees

We define a *stateful Polya tree* as a model extending the Polya tree, where each region in the recursive partition has one or more associated discrete state parameters with corresponding distributions over the possible states. A single stateful Polya tree implicitly generates a mixture of stateless Polya tree-like models, with the number of mixture components equal to the number of possible combinations over the

states of all regions in the tree, which may be exceedingly large. However, if states are either (i) independent between regions or (ii) have a simple dependency structure, such as Markov dependence between a region and its children in the recursive partition, we can do inference on the states and the Polya tree posterior without resorting to the representation as a mixture of Polya trees. Several of the Polya tree extensions discussed above, such as the two sample test in Holmes et al. (2015), can be expressed as stateful Polya trees with degenerate dependency structures between states. More flexible dependency structures can generate considerably more complex models, however.

The first extension of the Polya tree introducing state parameters for each region in the recursive partition was the Optional Polya tree of Wong and Ma (2010), where a state parameter is used to select the direction of partitioning, if any, in the recursive partitioning of a k-dimensional space. This model corresponds to a mixture of standard Polya trees, each with a different recursive partition, and the posterior could in principle be computed from that mixture. However, the states have a simple Markov dependence structure, and the authors show how a recursive algorithm allows computation of the posterior without resorting to the mixture representation. Such a recursive algorithm had previously been used for posterior computation of a wavelet model with states applied to each region in the wavelet expansion, and is closely related to algorithms for inference on linear Hidden Markov Models (Crouse et al., 1998).

Stateful Polya trees have also been used for Bayesian hypothesis testing in the two-sample problem (Ma and Wong, 2011; Soriano and Ma, 2017) and the multi-sample case, analogous to ANOVA (Ma and Soriano, 2016). These methods test a collection of local hypotheses rather than the single global hypothesis of Holmes et al. (2015). This gives them additional power to detect differences between samples which are localized in the sample space, for example a contamination in one part of

the sample space. In addition, examining the outcomes of the local test allows the researcher to characterize what part of the space, in particular, differentiates the samples.

Another recent development is the use of a state parameter to place a prior distribution on the Polya tree's concentration parameter (Ma, 2017). Difficulty dealing with the infinite-dimensional concentration parameter has led previous work to treat it as fixed at a value that ensures absolute continuity of the resulting distributions, or perhaps including a one-dimensional multiplicative factor that is treated as a tuning parameter (Berger and Guglielmi, 2001) or given a prior distribution (Hanson, 2006). Ma (2017) introduces a fully nonparametric prior for the concentration parameter using a state variable representing the magnitude of the parameter value on a given region, with Markov transitions from parent to child regions. The resulting model can be viewed as a mixture of standard Polya trees with different concentration parameters, but again recursion avoids dealing with the mixture representation for inference. We build on this model in Chapter 2 to create a hierarchy of Polya trees. In Chapter 3 we consider the advisability of using marginal likelihoods to determine model posterior probabilities, an important issue both in stateful Polya trees, where they are used to evaluate posterior probabilities of states, and in the broader Bayesian literature. Finally we conclude with some discussion of the work and other directions in which stateful Polya trees show untapped promise.

2

Hierarchical Adaptive Polya Trees

Many statistical problems involve learning from more than one sample of data. We may be interested in testing whether the distributions of two (or more) populations are statistically distinguishable from each other, or in learning more about a population by examining the distributions of a number of subpopulations. Classical approaches use low-dimensional parameterizations or summaries of the population distributions, as working with the infinite-dimensional distributions directly is challenging. Analysis of variance, for example, reduces distributions to a mean and variance, which are sufficient under the assumption of normality. A wide range of other parametric models within both the Bayesian and frequentist inferential frameworks use other parameterizations of the distribution to reduce the dimensionality of the problem. Classical nonparametric approaches use features of the samples such as medians (Westenberg, 1948), rank-based scores (Wilcoxon, 1945), or summaries of the empirical distribution functions, as in the Kolmogorov-Smirnov (Kolmogorov, 1933) and Cramér-von Mises tests (Anderson, 1962).

A number of Bayesian nonparametric approaches embrace the infinite-dimensional nature of the problem using extensions of Dirichlet processes. The Hierarchical

Dirichlet process of Teh et al. (2006) builds a hierarchical model using the Dirichlet process, allowing it to share information between samples. The Nested Dirichlet process of Rodríguez et al. (2008) takes a different approach, using a Dirichlet process as the base measure of a second Dirichlet process. This induces clustering in the samples, with samples in a cluster being modeled with a single density. Müller et al. (2004) model each sample density as a mixture of two components: one Dirichlet process mixture of Gaussians representing common structure between samples, and a second Dirichlet process mixture of Gaussians representing the idiosyncratic structure of the given sample. A variety of other dependent Dirichlet Processes (MacEachern, 1999) have been described in the literature. Teh and Jordan (2010) give an overview of hierarchical models based on the Dirichlet process.

While the Dirichlet process has been the basis of most of the work in this area, work has also been done on hierarchical extensions of other priors. For example, Teh (2006) defines a Hierarchical Pitman-Yor process, taking advantage of the more flexible clustering structure of the Pitman-Yor process over the Dirichlet process. Camerlenghi et al. (2017) consider hierarchical models based on Normalized Random Measures (Barrios et al., 2013), which includes both the Dirichlet process and the Pitman-Yor process as special cases.

The Polya tree has not previously been used to build hierarchical models for estimating and borrowing strength between multiple samples along the lines of the Hierarchical Dirichlet process or other discrete distribution processes, despite enjoying a number of advantages over the discrete processes. The Polya tree's infinite-dimensional concentration parameter allows much more flexibility for the modeling of variation between samples than the Dirichlet process (which has a single-dimensional concentration parameter) or other discrete processes. With an appropriate prior (Ma, 2017) on the concentration parameter we can learn a posterior distribution for the variation between sample distributions, and we show how to summarize that poste-

rior in an interpretable *variance function* that summarizes the amount of variation between densities at each point in the sample space. Also, because the Polya tree can model absolutely continuous distributions directly, we avoid the need for a mixture kernel, and the consequent necessity of MCMC methods for posterior inference. We describe such a model (Christensen and Ma, 2017) here, which we call the Hierarchical Adaptive Polya Tree (HAPT).

Following Ma (2017), HAPT uses states to borrow information spatially about the concentration parameters. One concentration parameter measures the shrinkage of the overall mean distribution to the prior measure, and is largely a nuisance parameter, unlikely to be of particular interest. The second concentration parameter, however, measures the variation of the sample distributions around the mean distribution, and may be a target of inferential interest in its own right. The infinite-dimensional concentration parameter measures variation at different locations and scales and is difficult to interpret directly beyond making qualitative comparisons between locations. We show how to transform it to a more interpretable summary of the variation between densities.

2.1 The Hierarchical Polya Tree

It is conceptually straightforward to extend the Polya tree to a hierarchical model. Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be k samples arising from related distributions on a complete, separable space Ω. For ease of exposition we again use $\Omega = (0, 1]$, though like the Polya tree the model can be applied to more general sample spaces. We model these samples as coming from k exchangeable distributions Q_i, which are centered at a common underlying measure Q, itself unknown. Applying Polya tree priors to both

Q and the Q_i gives us the hierarchical model

$$\begin{aligned} X_{ij} \mid Q_i &\overset{ind}{\sim} Q_i \\ Q_i \mid Q &\overset{iid}{\sim} \mathrm{PT}(Q, \boldsymbol{\tau}) \\ Q &\sim \mathrm{PT}(Q_0, \boldsymbol{\nu}). \end{aligned} \tag{2.1}$$

Here Q_0 is the overall prior mean, $\boldsymbol{\tau}$ controls the variation among samples around the common structure Q, and $\boldsymbol{\nu}$ controls the variation of the common structure from Q_0, which determines the smoothness of Q. This model, which we call the Hierarchical Polya tree, allows nonparametric estimation of the sample distributions Q_i and the common structure Q. The Hierarchical Polya tree model is illustrated in Figure 2.1. The first row shows the upper level of the hierarchy, which like the basic Polya tree illustrated in Figure 1.1 is centered at the uniform distribution on $(0, 1]$. The second row shows the lower level of the hierarchy, the individual group distributions Q_i conditioned on Q. They follow exactly the same Polya tree construction, but each cut is centered on the corresponding cut from Q, rather than on the uniformed distribution. Q captures the common structure of the groups, while the Q_i model the idiosyncratic structure of each group.

The hierarchy of Polya trees translates directly to the decomposed space as a hierarchical model for Beta random variables. For an arbitrary region $A \in \mathcal{A}$, we have

$$\begin{aligned} \theta_i(A) \mid \theta(A) &\overset{iid}{\sim} \mathrm{Beta}(\theta(A)\tau(A), (1-\theta(A))\tau(A)) \\ \theta(A) &\sim \mathrm{Beta}(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A)). \end{aligned} \tag{2.2}$$

The representation of the hierarchy of Polya trees as a hierarchy of Beta random variables allows tractable posterior inference, as described in Section 2.5.

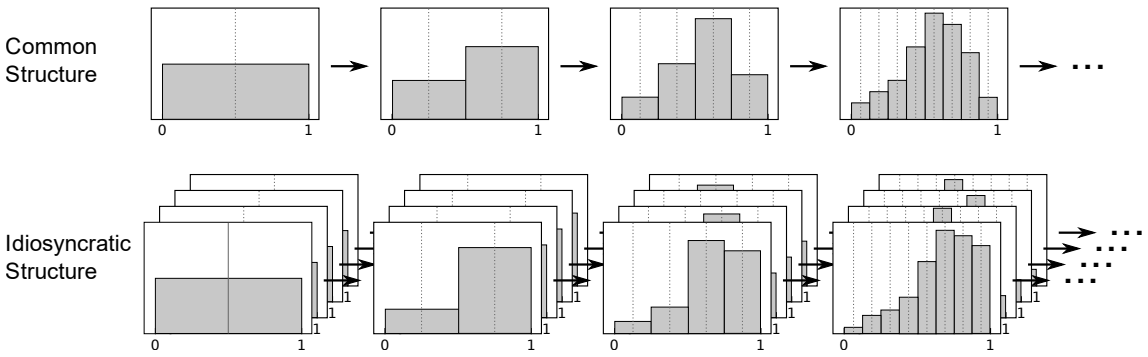

FIGURE 2.1: An illustration of the Hierarchical Polya tree.

2.2 The Stochastically Increasing Shrinkage prior on dispersion

The Polya tree's concentration parameter is traditionally set to increase with depth at a predetermined rate to ensure absolute continuity, with a constant multiplicative term to control the overall level of variation which may be treated as a tuning parameter (as in Berger and Guglielmi (2001)) or have a prior placed on it. Hanson (2006) discusses some of the necessary considerations when placing a prior on this parameter. However, Ma (2017) shows that putting a fully nonparametric prior on the concentration parameter allows the Polya tree to learn the true distribution more accurately, particularly when the smoothness of the underlying density varies over the sample space. We can extend the Hierarchical Polya tree model by placing priors on both concentration parameters. In addition to more accurate inference on Q and the Q_i, putting a prior distribution on $\boldsymbol{\tau}$ allows us to learn the variation between samples in a nonparametric way. That is, we can estimate a posterior *dispersion function* which summarizes the variability between sample densities at any given point in the sample space. Dispersion can be measured in a variety of ways; in Section 2.7.1 we show how to derive the posterior mean variance of the densities and interpret a standardized version using the coefficient of variation to correct for the height of the density. Nonparametric inference on the variation between samples across the sample space is made possible by the flexibility of the Polya tree model. We contrast how several other models treat between-sample variation in Section 2.4.

A simple approach is to put independent priors on the variance of each Beta random variable. However, we expect spatial structure in the dispersion function—depending on how smooth the underlying densities are, the amount of Beta variance is generally smaller for deeper levels of the partition though the decay in the Beta variance can be heterogeneous over the sample space depending on the local smoothness of the densities. While the recursive partitioning allows independent priors to

capture some spatial structure, we can do better by introducing dependency between regions in the partition. Ma (2017) introduces Markov dependency on the concentration parameter, following the tree topology. The Stochastically Increasing Shrinkage (SIS) prior introduces a state variable $S(A)$ supported on a finite set of integers $1, \ldots, I$, corresponding to decreasing prior variance and increasing shrinkage for the Beta random variables. For example, we may have $S(A) \in \{1, 2, 3, 4\}$ with $S(A) = i$ implying $\nu(A) \sim F_i$ with the F_i stochastically ordered $F_1 \prec F_2 \prec F_3 \prec F_4$ and F_4 being a point mass at infinity. The number of states and the corresponding distributions can be chosen to balance the flexibility and computational complexity of the model. A simple way to enforce such a stochastic ordering is through partitioning the support of the concentration parameter $\nu(A)$ into disjoint intervals. Given these latent states, the SIS prior adopts a transition probability matrix for $S(A) \mid S(\text{Par}(A))$, denoted $\Gamma(A)$. Ma (2017) discusses several prior possibilities for this transition matrix, including an empirical Bayes estimate. We adopt the simplest recommendation there that provides generally adequate flexibility, namely

$$\Gamma(A) = \begin{bmatrix} \frac{1}{I} & \frac{1}{I} & \cdots & \frac{1}{I} \\ 0 & \frac{1}{I-1} & \cdots & \frac{1}{I-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This upper-triangular transition probability matrix induces stochastically increasing shrinkage, and ensures that the model generates absolutely continuous densities (see Theorem 1). We denote the SIS prior

$$\boldsymbol{\nu} \sim \text{SIS}(\boldsymbol{\Gamma}).$$

Figure 2.2 illustrates the SIS prior in action. As you move down to finer resolutions the shrinkage state tends to increase, eventually reaching complete shrinkage. Because the transitions are stochastic and the state of each node can be learned from

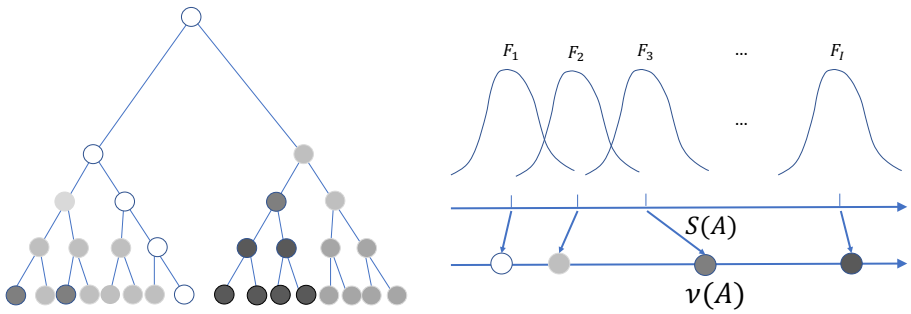

FIGURE 2.2: Illustration of the SIS prior. The shrinkage states increase as you follow the tree down to finer scales, eventually reaching complete shrinkage—but allowing less shrinkage where the data dictates such.

the data, the amount of shrinkage or prior variance can vary over the sample space to capture large smooth features in one part of the sample space and smaller scale features in another part. This allows the resulting density to have heterogeneous smoothness across the sample space.

2.3 The Hierarchical Adaptive Polya Tree

Having described the SIS prior, we can adopt this prior for the concentration parameters $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$, and write the complete model as follows:

$$
\begin{aligned}
X_{ij} \mid Q_i &\overset{ind}{\sim} Q_i \\
Q_i \mid Q, \boldsymbol{\tau} &\overset{iid}{\sim} \mathrm{PT}(Q, \boldsymbol{\tau}) \\
Q \mid \boldsymbol{\nu} &\sim \mathrm{PT}(Q_0, \boldsymbol{\nu}) \\
\boldsymbol{\tau} &\sim \mathrm{SIS}(\boldsymbol{\Gamma}_{\boldsymbol{\tau}}) \\
\boldsymbol{\nu} &\sim \mathrm{SIS}(\boldsymbol{\Gamma}_{\boldsymbol{\nu}}).
\end{aligned}
$$

We call this model the Hierarchical Adaptive Polya Tree (HAPT). In contrast to existing models, this specification allows fully nonparametric inference on both the densities and the variation between densities. The inclusion of the SIS priors also allows the model to adapt the level of shrinkage or information borrowing in different parts of the sample space to more accurately capture the densities of each group, rather than using fixed uniform shrinkage. Figure 2.3 shows a graphical representation of the model.

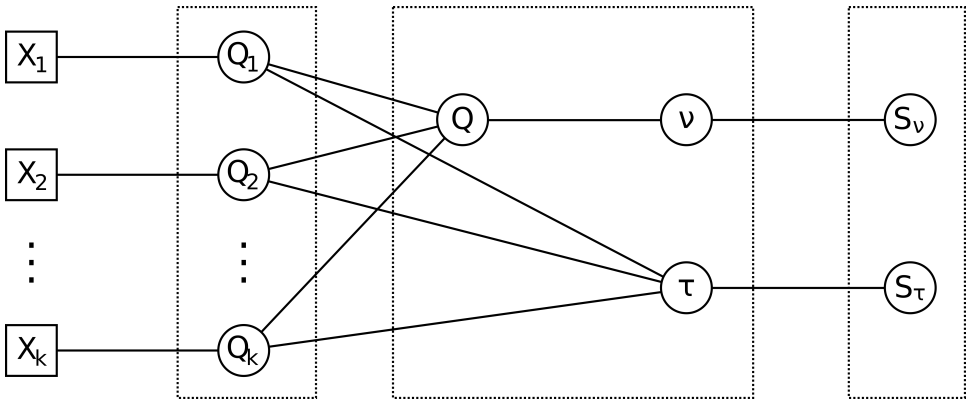

FIGURE 2.3: A graphical representation of the model. The boxes outline the parts of the model whose posteriors are computed with each of the three strategies described in Section 2.5.2. From left to right: The posterior of the Q_i conditional on other parameters is conjugate and can be integrated out numerically; the posterior of $Q, \boldsymbol{\nu}$, and $\boldsymbol{\tau}$ conditional on $\boldsymbol{S_\nu}$ and $\boldsymbol{S_\tau}$ is approximated using quadrature; and the posterior of $\boldsymbol{S_\nu}$ and $\boldsymbol{S_\tau}$ is computed using HMM methods.

2.4 Comparison to existing models

The most prominant existing nonparametric models for estimation of related distributions are based on the Dirichlet process, including the Hierarchical Dirichlet process (HDP) (Teh et al., 2006), Nested Dirichlet process (NDP) (Rodríguez et al., 2008), and the Hierarchichal Dirichlet Process Mixture (HDPM) (Müller et al., 2004). The Hierarchical Adaptive Polya tree enjoys several advantages over these methods:

1. **Nonparametric estimation of between-sample variation.** The HDP and NDP have a single concentration parameter that controls the dependence between samples. The HDPM has one parameter per sample that controls what proportion of the sample is explained by common structure and how much by idiosyncratic structure.

   In contrast, HAPT places a fully nonparametric prior on the variation between samples, which allows it to learn spatially heterogeneous variation between samples. Indeed, the variation between samples at different locations of the sample space may be of primary interest in some scientific applications: learning where common structure is largely preserved between samples and where distributions vary widely may point the way to understanding important underlying phenomena.

2. **Computation.** HDP, NDP, and HDPM all rely on MCMC methods to draw from the posterior. The Hierarchical Adaptive Polya tree is not fully conjugate, but the necessary integration can be split into low-dimensional integrals and approximated extremely quickly using adaptive quadrature methods, without concerns about Markov chain convergence. See Section 2.5.2 for details.

3. **Interpretability.** HAPT provides an easily interpretable estimate of the common structure: The posterior estimate of Q is both the estimate of the mean

density across samples, and the expected value of the posterior predictive for a new sample. In contrast, the HDP estimates a discrete instead of continuous distribution, and the NDP does not provide any estimate of common structure. The HDPM provides an estimate of common structure, but it is neither the mean of sample distributions nor a posterior predictive estimate. Interpretation of the common structure in the HDPM model is most straightforward if variation between samples involves contamination of an underlying distribution by an idiosyncratic process for each sample.

Other existing models based on Normalized Random Measures (Camerlenghi et al., 2017), including the Hierarchical Pitman-Yor model (Teh, 2006), suffer from drawbacks similar to those of the models based on the Dirichlet process, including not allowing flexible modeling of the dispersion between samples, requiring MCMC samplers, and requiring mixture kernels to model continuous distributions.

2.5 Bayesian inference and computation

The HAPT model is partially conjugate: the conditional posterior for $Q_i \mid Q, \boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{S_\tau}, \boldsymbol{S_\nu}$ is a standard Polya tree. Though not conjugate, the remaining joint posterior for $Q, \boldsymbol{\tau}, \boldsymbol{\nu} \mid \boldsymbol{S_\tau}, \boldsymbol{S_\nu}$ can be reliably approximated using adaptive quadrature methods. The computational strategies used are described in Section 2.5.2

To derive the posterior we use the representation of the Polya trees Q and Q_i in terms of Beta-distributed random variables $\theta(A)$ and $\theta_i(A)$ for each node A of the tree. With this notation, The second and third lines in (2.1) can be written in terms of the θ and θ_i as in Equation 2.2:

$$\theta_i(A) \mid \theta(A) \overset{iid}{\sim} \text{Beta}\left(\theta(A)\tau(A), (1-\theta(A))\tau(A)\right)$$

$$\theta(A) \sim \text{Beta}\left(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A)\right).$$

Including the concentration parameters, we can write the posterior for the parameters of a region A in the following form conditional on the state parameters $S_\tau(A), S_\nu(A)$:

$$\begin{aligned}\pi(\theta(A), \tau(A), \nu(A) \mid S_\tau(A), S_\nu(A), \boldsymbol{X}) \propto & \\ & \theta(A)^{\theta_0(A)\nu(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu(A)-1} \times \\ & \left[B\left(\theta(A)\tau(A), (1-\theta(A))\tau(A)\right)\right]^{-k} \times \\ & \prod_{i=1}^{k} B(\theta(A)\tau(A) + n_i(A_\ell), (1-\theta(A))\tau(A) + n_i(A_r)) \times \\ & \pi(\tau(A) \mid S_\tau(A))\pi(\nu(A) \mid S_\nu(A)).\end{aligned} \tag{2.3}$$

where $B(\cdot, \cdot)$ is the Beta function. The full posterior is the summation of Equation (2.3) over the possible states of $\boldsymbol{S_\tau}$ and $\boldsymbol{S_\nu}$, with their respective priors factored in.

2.5.1 Derivation of the posterior

We use the following notation in this section:

- A is a set in the recursive partition of Ω implicit in the Polya tree.
- A_ℓ and A_r indicate the left and right children of A, respectively; these are also sets in the recursive partition of Ω.
- $n(A)$ is the number of data points across all samples that are contained in A; $n_i(A)$ is the number of datapoints in sample i that are contained in A. Thus $n(A) = \sum_i n_i(A)$.

We ignore for the time being the priors on $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$, as they are not important for this derivation and can be reinserted at the end.

The density of the distribution generated from the HAPT model at an observation x_{ij} consists of three factors:

1. The baseline density $q_0(x_{ij})$

2. A term indicating how the common structure modifies the density:

$$\prod_{A:x_{ij}\in A_\ell} \theta(A)\frac{Q_0(A)}{Q_0(A_\ell)} \prod_{A:x_{ij}\in A_r} (1-\theta(A))\frac{Q_0(A)}{Q_0(A_r)}$$

   Note that under the canonical representation, $\frac{Q_0(A)}{Q_0(A_\ell)} \equiv 2$

3. A term indicating how the idiosyncratic structure of the particular sample containing x modifies the density:

$$\prod_{A:x_{ij}\in A_\ell} \theta_i(A)\frac{Q(A)}{Q(A_\ell)} \prod_{A:x_{ij}\in A_r} (1-\theta_i(A))\frac{Q(A)}{Q(A_r)}$$

   Note that by definition, $\frac{Q(A)}{Q(A_\ell)} = 1/\theta(A)$ and $\frac{Q(A)}{Q(A_r)} = 1/(1-\theta)$.

Altogether this gives us the following likelihood:

$$f(\mathbf{X} \mid \theta, \theta_i) = \prod_{x_{ij}} \left[q_0(x_{ij}) \prod_{A:x_{ij}\in A_\ell} \left[\theta(A)\frac{Q_0(A)}{Q_0(A_\ell)}\right] \prod_{A:x_{ij}\in A_r} \left[(1-\theta(A))\frac{Q_0(A)}{Q_0(A_r)}\right] \right.$$
$$\left. \prod_{A:x_{ij}\in A_\ell} \left[\theta_i(A)\frac{Q(A)}{Q(A_\ell)}\right] \prod_{A:x_{ij}\in A_r} \left[(1-\theta_i(A))\frac{Q(A)}{Q(A_r)}\right] \right]$$
$$= \prod_{x_{ij}} \left[q_0(x_{ij}) \prod_{A:x_{ij}\in A_\ell} \left[\theta_i(A)\frac{Q_0(A)}{Q_0(A_\ell)}\right] \prod_{A:x_{ij}\in A_r} \left[(1-\theta_i(A))\frac{Q_0(A)}{Q_0(A_r)}\right] \right].$$

Rearranging terms, the likelihood can be written as follows:

$$f(\mathbf{X} \mid \theta_i, \theta) = \prod_{x_{ij}} q_0(x_{ij}) \times$$
$$\prod_A \left[\prod_{i=1}^{k} \left[\theta_i(A)\frac{Q_0(A)}{Q_0(A_\ell)}\right]^{n_i(A_\ell)} \left[(1-\theta_i(A))\frac{Q_0(A)}{Q_0(A_r)}\right]^{n_i(A_r)} \right],$$

which gives the following form for the posterior of $\theta(A), \theta_i(A)$ for a particular A:

$$\pi(\theta(A), \theta_i(A) \mid \tau(A), \nu(A), x) \propto \theta(A)^{\theta_0(A)\nu(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu(A)-1} \times \left[\frac{\Gamma(\tau(A))}{\Gamma(\theta(A)\tau(A))\Gamma((1-\theta(A))\tau(A))}\right]^k \times \prod_{i=1}^{k}\left[\theta_i(A)^{\theta(A)\tau(A)+n_i(A_\ell)-1}(1-\theta_i(A))^{(1-\theta(A))\tau(A)+n_i(A_r)-1}\right].$$

Conditional on $\theta(A)$ and $\tau(A)$ the $\theta_i(A)$ are Beta distributed, and we can integrate them out analytically:

$$\pi(\theta(A) \mid \tau(A), \nu(A), x) \propto \theta(A)^{\theta_0(A)\nu(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu(A)-1} \times \left[\frac{\Gamma(\tau(A))}{\Gamma(\theta(A)\tau(A))\Gamma((1-\theta(A))\tau(A))}\right]^k \times \prod_{i=1}^{k}\frac{\Gamma\left(\theta(A)\tau(A)+n_i(A_\ell)\right)\Gamma\left((1-\theta(A))\tau(A)+n_i(A_r)\right)}{\Gamma\left(\tau(A)+n_i(A)\right)}.$$

We can then simply multiply by the priors on $\tau(A)$ and $\nu(A)$ to obtain

$$\pi(\theta(A), \tau(A), \nu(A) \mid S_\tau(A), S_\nu(A), x) \propto \theta(A)^{\theta_0(A)\nu(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu(A)-1} \times \left[\frac{\Gamma(\tau(A))}{\Gamma(\theta(A)\tau(A))\Gamma((1-\theta(A))\tau(A))}\right]^k \times \prod_{i=1}^{k}\frac{\Gamma\left(\theta(A)\tau(A)+n_i(A_\ell)\right)\Gamma\left((1-\theta(A))\tau(A)+n_i(A_r)\right)}{\Gamma\left(\tau(A)+n_i(A)\right)} \times \pi(\tau(A))\pi(\nu(A)).$$

2.5.2 Computation

Posterior computation of the HAPT model requires three distinct computational strategies. We split the model (see Figure 2.3) into three parts, each of which requires a different approach. Each part of the model is conditioned on all parameters which are further to the right in Figure 2.3. We describe how to integrate out each of the

first two parts, which reduces the problem to evaluating the posterior probabilities of all possible combinations of the state variables $\boldsymbol{S}_{\boldsymbol{\tau}}, \boldsymbol{S}_{\boldsymbol{\nu}}$.

1. $\pi(Q_i \mid Q, \boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{S}_{\boldsymbol{\tau}}, \boldsymbol{S}_{\boldsymbol{\nu}}, \boldsymbol{X})$: The individual sample densities Q_i, conditional on all other parameters, are *a priori* distributed according to a standard Polya tree. The corresponding conditional posterior is therefore also a Polya tree. This allows us to integrate out the Q_i when computing the posterior. If individual sample densities are of inferential interest their posteriors can easily be reconstructed after the main posterior computation is completed.

2. $\pi(Q, \boldsymbol{\tau}, \boldsymbol{\nu} \mid \boldsymbol{S}_{\boldsymbol{\tau}}, \boldsymbol{S}_{\boldsymbol{\nu}}, \boldsymbol{X})$: The remaining continuous parts of the joint model, namely the common structure Q and the continuous concentration parameters $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$, conditioned on the discrete state parameters $\boldsymbol{S}_{\boldsymbol{\tau}}$ and $\boldsymbol{S}_{\boldsymbol{\nu}}$, are not conjugate and must be integrated numerically. Because all parameter dependence across nodes in the Polya tree topology is restricted to the discrete state parameters, by conditioning on those parameters we are able to compute the posterior of the remaining parameters for each node of the tree independently.

   This has two significant implications. First, rather than tackling a very high dimensional integral over the product space of the parameters for all nodes, we have a much more tractable collection of low-dimensional integrals: we need only integrate the three-dimensional joint posterior of $\theta(A), \tau(A), \nu(A)$ for each region A in the Polya tree. Each of these integrals is tractable using adaptive quadrature techniques. Second, these integrals can be computed in parallel.

   An additional observation allows us to further accelerate the adaptive quadrature. We note that the joint posterior distribution for $\theta(A), \tau(A), \nu(A)$ conditional on $S_\tau(A)$ and $S_\nu(A)$, with the other parameters integrated out, can be

factored as

$$\pi(\theta(A), \tau(A), \nu(A) \mid S_\nu(A), S_\tau(A), \boldsymbol{X}) = g(\theta(A), \tau(A)) \times h(\theta(A), \nu(A)).$$

This allows us to factor the three dimensional integral:

$$\begin{aligned}\iiint \pi(\theta(A), \tau(A), \nu(A) \mid S_\nu(A), S_\tau(A), x)\, d\tau(A)\, d\nu(A)\, d\theta(A) \\ = \int \left[\int g(\theta(A), \tau(A))\, d\tau(A)\right] \left[\int h(\theta(A), \nu(A))\, d\nu(A)\right] d\theta(A),\end{aligned}$$

This factorization effectively reduces the dimensionality of the integral: rather than evaluating the unnormalized posterior at points throughout a 3-dimensional space, we need only evaluate it on the union of two 2-dimensional spaces.

3. The posterior distribution of the state parameters $\boldsymbol{S_\tau}$ and $\boldsymbol{S_\nu}$ at first appears to be the most intimidating part of the model: It is a distribution over the product space of a large number of discrete parameters, resulting in an enormous number of level combinations. The naive computation of the joint posterior,

$$P\left(\bigcap_{A \in \mathcal{A}} S_\tau(A) = i_A, S_\nu(A) = j_A\right)$$

is straightforward but needs to be repeated for every possible combination of $S_\tau(A)$ and $S_\nu(A)$ for every node in the tree, which is computationally prohibitive. Here the Markov dependency structure comes to our rescue. The shrinkage states constitute a Hidden Markov Model following the tree topology (Crouse et al., 1998), and we can factor the joint distribution and calculate the posterior probabilities using a forward-backward algorithm in a manner analogous to inference strategies for linear Hidden Markov Models.

During the forward-backward algorithm we can compute expectations of any

function that can be expressed in the form

$$f(\cdot) = \prod_A f^*(\theta(A), \tau(A), \nu(A)),$$

where $f^*(\cdot)$ is an arbitrary function in L_1. This includes the marginal likelihood, the expected value of the estimated common density $q(\cdot)$ or any individual sample density $q_i(\cdot)$ at any given point, expectations of random variables $Y \sim Q$ or $Y_i \sim Q_i$, and a wide variety of other functions, such as the variance function described in Section 2.7.1.

This computation is recursive, and we give a brief example of how it is carried out for the marginal likelihood. The previous two computational strategies give us the ability to calculate (up to quadrature approximation) the marginal likelihood, within a given node, of all remaining parameters conditional on $\boldsymbol{S_\tau}$ and $\boldsymbol{S_\nu}$. Combining this with the prior transition parameters specified in $\boldsymbol{\Gamma}$ we are able to calculate the posterior probabilities for $S_\tau(A)$ and $S_\nu(A)$, and the overall marginal likelihood of the distribution on A, by considering the recursively-calculated marginal likelihoods of the child regions of A under each possible state.

Obviously this recursion requires a stopping point. The simplest method is to truncate the tree at a predetermined depth. Hanson (2006) offers some guidance on how to choose the depth of a truncated Polya tree based on sample size and other considerations. We recommend using as large a tree as is computationally feasible in order to minimize approximation errors due to truncation. If the data deviate very strongly from the prior distribution—as is common, for example, in high-dimensional settings—a more sophisticated approach may be required, such as truncating a branch of the tree when it reaches a depth where the node contains only a few data points.

2.6 Theoretical results

Theorem 1. *(Absolute Continuity) Let $Q, Q_1, \ldots, Q_k$ be random measures distributed according to a HAPT model with base measure Q_0. If the SIS priors on $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$ each include a complete shrinkage state that absorbs all possible chains in a finite number of steps with probability 1, then $Q, Q_1, \ldots, Q_k \ll Q_0$ with probability 1.*

Remark 1. A sufficient condition for the complete shrinkage state to absorb all chains in a finite number of steps with probability 1 is that the transition probability from every other state to the complete shrinkage state is bounded away from zero, which is satisfied by our choice of $\boldsymbol{\Gamma}$.

Proof. The result follows directly from repeated application of Theorem 3 in Ma (2017). From one application of that theorem we have that $Q \ll Q_0$ with probability 1; by a second application we have that $Q_i \ll Q$ for each Q_i with probability 1. Thus $Q_i \ll Q_0$ with probability 1. □

Theorem 2. *(Prior support) Let $f, f_1, \ldots, f_k$ be the probability density functions corresponding to arbitrary absolutely continuous distributions on Ω. Let $q, q_1, \ldots, q_k$ be corresponding densities from a HAPT model satisfying*

1. *Q_0 and μ are equivalent measures, that is $Q_0 \ll \mu$ and $\mu \ll Q_0$;*
2. *$I_{\boldsymbol{\tau}}, I_{\boldsymbol{\nu}} \geqslant 2$, i.e. there are at least two shrinkage states at each level.*

Then $f, f_1, \ldots, f_k$ are in the L_1 prior support.

Proof. The upper level of the hierarchy (q) follows immediately from Theorem 4 in Ma (2017), since its prior coverage is not altered by the addition of the hierarchy. For the second level of the hierarchy, note that by Theorem 4 in Ma (2017) we have

$$\forall q \ll \mu,\ P\Big(\int |q_i(x) - f_i(x)| d\mu < \tau_i \mid q\Big) > 0.$$

Further, by definition of conditional expectation,

$$P\left(\int |q_i(x) - f_i(x)|d\mu < \tau_i \mid q\right) = E\left(\mathbf{1}\left[\int |q_i(x) - f_i(x)|d\mu < \tau_i\right] \mid q\right)$$

is measurable with respect to Q. Thus

$$P\left(\int |q_i(x) - f_i(x)|d\mu < \tau_i\right) = \int P\left(\int |q_i(x) - f_i(x)|d\mu < \tau_i \mid q\right)\, dQ > 0,$$

as the integrand is positive with probability 1. □

Our final result gives posterior consistency in the case where the groups have equal sample sizes:

Theorem 3. *(Posterior consistency) Let $\boldsymbol{D}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$ be observed data consisting of k independent samples, each of size n, from absolutely continuous distributions $P_1, \ldots, P_k$. Let $\pi(\cdot)$ be a Hierarchical Adaptive Polya tree model on the k densities with overall prior mean Q_0^*, and let $\pi(\cdot \mid \boldsymbol{D}_n)$ be the corresponding posterior. If $P_i \ll Q_0^*$ for all i then we have posterior consistency under the weak topology. That is,*

$$\pi(U \mid \boldsymbol{D}_n) \to 1 \text{ as } n \to \infty$$

for any weak neighborhood U of the product measure $P_1 \times \cdots \times P_k$.

Proof. By Schwartz's theorem, it is sufficient to show that $P_1, \ldots, P_k$ jointly lie in the Kullback-Leibler support of the prior. The proof proceeds as follows: We consider the product space of all k distributions so that we can show joint convergence. We restrict ourselves to a compact set with mass $1 - \epsilon'$, and define a set $\tilde{D}_\epsilon$, which depends on p_0. We show that $\tilde{D}_\epsilon$ has positive prior mass. We then show that by choosing ϵ and ϵ' appropriately, we can make $\tilde{D}_\epsilon$ lie within an arbitrarily small K-L ball around p_0. This shows that $P_1, \ldots, P_k$ are jointly in the support of the prior, and concludes the proof. We follow closely the proof of Theorem 5 in Ma (2017).

Let P_0 be the product measure $P_0 = P_1 \otimes \cdots \otimes P_k$ on Ω^k, and let $p_0 = dP_0/d\mu$ be the corresponding density. Let $Q_0 = Q_0^* \otimes \cdots \otimes Q_0^*$, and let $q_0 = dQ_0/d\mu$. Define additionally the densities $\tilde{p}_0 = dP_0/dQ_0$ and $\tilde{q} = dQ/dQ_0$ for any $Q \ll Q_0$. Let M be a finite upper bound on $\tilde{p}_0$. The Kullback-Leibler distance between p_0 and q is given by

$$\mathrm{KL}_\mu(p_0, q) = \int p_0 \log(p_0/q)\, d\mu = \int \tilde{p}_0 \log(\tilde{p}_0/\tilde{q})\, dQ_0 = \mathrm{KL}_{Q_0}(\tilde{p}_0, \tilde{q}).$$

By Lusin's theorem, for any $\epsilon' > 0$ there exists a compact set $E \subset \Omega$ with $Q_0(E^c) < \epsilon'$, such that $\tilde{p}_0$ is continuous (and so uniformly continuous) on E. For any $\epsilon > 0$ there exists a partition $\Omega = \cup_i A_i$ with all $A_i \in \mathcal{A}^(k)$ for some k, such that the diameter of each $A_i \cap E$ is less than ϵ. We define

$$\delta_E(\epsilon) = \sup_{x,y \in E : |x-y| < \epsilon} |\tilde{p}_0(x) - \tilde{p}_0(y)|$$

and

$$d_i = \max\left(\sup_{A_i \cap E} \tilde{p}_0(x) + \delta_E(\epsilon), \epsilon' \right).$$

Note that because p_0 is uniformly continuous on E, $\delta_E(\epsilon) \to 0$ as $\epsilon \to 0$.

Let $D_\epsilon(\tilde{p}_0)$ be the collection of step functions $g(x) = \sum_i g_i \mathbf{1}_{A_i}(x)$ with $d_i \leqslant g_i < d_i + \delta_E(\epsilon)$. For every $g \in D_\epsilon(\tilde{p}_0)$, let $\tilde{g} = g / \int g\, dQ_0$ be the normalized version of g, and let $\tilde{D}_\epsilon(\tilde{p}_0)$ be the collection of the $\tilde{g}$.

Let I be the number of sets A_i in the partition $\Omega = \cup_i A_i$. We can consider each step function g as a point in I-dimensional space, where the ith dimension corresponds to the value of the step function on the set A_i. Note that $D_\epsilon(\tilde{p}_0) = [d_1, d_1 + \delta_E(\epsilon)) \times \cdots \times [d_I, d_I + \delta_E(\epsilon))$ is a convex set in this I-dimensional space, and we can find an open ball in this I-dimensional space which is a subset of $D_\epsilon(\tilde{p}_0)$. The normalized version $\tilde{D}_\epsilon(\tilde{p}_0)$ is in turn a convex set in the $(I-1)$-simplex, which also contains an open ball in the $(I-1)$-simplex.

The HAPT model places positive probability on normalized step functions taking unique values precisely on the sets A_i, i.e. the same $(I-1)$-simplex noted above. Because each of the mass assignment parameters in HAPT has a Beta prior, HAPT has positive prior density everywhere in the $(I-1)$ simplex. Because $\tilde{D}_\epsilon(\tilde{p}_0)$ contains an open ball, it follows that it has positive prior mass.

It now remains to show that we can bound $\tilde{D}_\epsilon(\tilde{p}_0)$ within an arbitrarily small ball about p_0. The remainder of the proof follows exactly as Ma (2017). For every $\tilde{g} \in \tilde{D}_\epsilon(\tilde{p}_0)$, we have

$$\begin{aligned} 0 \leqslant \mathrm{KL}_{Q_0}(\tilde{p}_0, \tilde{g}) &= \int \tilde{p}_0 \log(\tilde{p}_0/\tilde{g})\, dQ_0 \\ &= \int \tilde{p}_0 \log(\tilde{p}_0/g)\, dQ_0 + \log\left(\int g\, dQ_0\right) \\ &= \int_E \tilde{p}_0 \log(\tilde{p}_0/g)\, dQ_0 + \int_{E^c} \tilde{p}_0 \log(\tilde{p}_0/g)\, dQ_0 + \log\left(\int g\, dQ_0\right). \end{aligned}$$

Because $g \geqslant p_0$ everywhere in E, the first integral is not greater than 0. The second integral is bounded by $M \log(M/\epsilon')\epsilon'$, which goes to zero as $\epsilon' \to 0$. To bound the third term, we note that

$$\log\left(\int g\, dQ_0\right) = \log\left(1 + \int (g - \tilde{p}_0)\, dQ_0\right) \leqslant \int (g - \tilde{p}_0)\, dQ_0,$$

We can bound this last integral by

$$\int (g - \tilde{p}_0)\, dQ_0 \leqslant \int_E (g - \tilde{p}_0)\, dQ_0 + \int_{E^c} |g - \tilde{p}_0|\, dQ_0.$$

On the set E, we have $g_0 - \tilde{p}_0 \leqslant 3\delta_E(\epsilon) + \epsilon'$, and on E^c we have $g_0 - \tilde{p}_0 \leqslant 3M + \epsilon'$. Thus

$$\int (g - \tilde{p}_0)\, dQ_0 \leqslant 3\delta_E(\epsilon) + \epsilon' + (3M + \epsilon')\epsilon' \to 0 \text{ as } \epsilon, \epsilon' \to 0.$$

This shows that by picking ϵ, ϵ' appropriately, $\tilde{D}_\epsilon(\tilde{p}_0)$ is contained within an arbitrarily small KL ball about $\tilde{p}_0$, and so p_0 is in the KL support of the HAPT model. $\square$

2.7 Methodological applications of HAPT

We present two ways in which the HAPT model can be applied to infer structure that existing models have not been able to capture. The first application is the ability of HAPT to model the "dispersion function" (defined below) on the sample space; we show how to calculate the fitted dispersion function from the $\boldsymbol{\tau}$ parameter. No existing model permits inference on the variation between sample densities in this manner. The second application is clustering samples based on their distributions, while allowing for within-cluster variation. While the Nested Dirichlet process clusters distributions, it allows no variation (beyond sampling variation) within clusters. The importance of allowing for variation within clusters was first pointed out by MacEachern (2008), who described a dependent Dirichlet process which would incorporate within-cluster variation.

2.7.1 Inferring the between-sample dispersion function

The primary target of inference in problems with multiple samples is often the variation between samples. It is this inference, for example, which lends ANOVA its name, though the model is typically presented in terms of the overall and sample means.

In the HAPT model, variation between samples is captured by $\boldsymbol{\tau}$, an infinite-dimensional parameter which estimates variation at different locations and scales. Rather than trying to provide guidance on how to interpret the multiscale structure in $\boldsymbol{\tau}$, we show how to recast it into an estimate of the variation between samples at any given point in the sample space, giving us a posterior *dispersion function* analogous to the posterior mean function.

Let $q_\star$ be the posterior predictive density for a new sample, with corresponding Beta-distributed random variables $\theta_\star(A)$ for each region A in the recursive partition.

Since $q_\star$ is random, we can estimate the variance function $v : \Omega \to \mathbb{R}^+$ which gives, for any point x in the sample space, the variance of $q_\star(x)$ conditional on the density of the common structure, $q(\cdot)$. This is analogous to the variance between treatments in traditional ANOVA. We have

$$q_*(x) = q_0(x) \prod_{A \ni x} ||A|| \left(\frac{\theta_\star(A)}{||A_\ell||} \right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1 - \theta_\star(A)}{||A_r||} \right)^{\mathbf{1}(x \in A_r)},$$

where $q_0(\cdot)$ is the density corresponding to the prior mean distribution Q_0. We note that

$$\theta_\star(A) | \theta(A), \tau(A) \sim \text{Beta}(\theta(A)\tau(A), (1 - \theta(A))\tau(A)).$$

We will need the facts that

$$\mathbb{E}(\theta_\star(A) \mid \theta(A), \tau(A)) = \theta(A)$$

and

$$\text{Var}(\theta_\star(A) \mid \theta(A), \tau(A)) = \text{Var}(1 - \theta_\star(A) \mid \theta(A), \tau(A)) = \frac{\theta(A)(1 - \theta(A))}{\tau(A) + 1},$$

which together imply

$$\begin{aligned}
\mathbb{E}(\theta_\star(A)^2 \mid \theta(A), \tau(A)) &= \frac{\theta(A)(1 - \theta(A))}{\tau(A) + 1} + \theta(A)^2 \\
&= \frac{\theta(A)(\theta(A)\tau(A) + 1)}{\tau(A) + 1} \\
\mathbb{E}((1 - \theta_\star(A))^2 \mid \theta(A), \tau(A)) &= \frac{\theta(A)(1 - \theta(A))}{\tau(A) + 1} + (1 - \theta(A))^2 \\
&= \frac{(1 - \theta(A))((1 - \theta(A))\tau(A) + 1)}{\tau(A) + 1}.
\end{aligned}$$

Let $\boldsymbol{\theta}(x) = \{\theta(A) : A \ni x\}$. It is clear that $\boldsymbol{\theta}(x)$ contains exactly the information about q which is relevant to $q_\star(x)$; that is, the distribution of $q_\star(x) \mid \boldsymbol{\theta}(x)$ is identical

to the distribution of $q_\star(x) \mid q$. With the above, we can calculate

$$\mathrm{Var}\left(q_\star(x) \mid q\right) = \mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\mathrm{Var}\left(q_\star(x) \mid q, \boldsymbol{\tau}, \mathbf{S}_{\boldsymbol{\tau}}\right) \mid q\right] + \mathrm{Var}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\mathbb{E}\left(q_\star(x) \mid q, \boldsymbol{\tau}, \mathbf{s}_{\boldsymbol{\tau}}\right) \mid q\right] =$$

$$q_0(x)^2 \mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\mathrm{Var}\left(\prod_{A \ni x} ||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)} \middle| \boldsymbol{\theta}(x), \boldsymbol{\tau}, \mathbf{S}_{\boldsymbol{\tau}}\right) \middle| \boldsymbol{\theta}(x)\right] +$$

$$q_0(x)^2 \mathrm{Var}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\mathbb{E}\left(\prod_{A \ni x} ||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)} \middle| \boldsymbol{\theta}(x), \boldsymbol{\tau}, \mathbf{S}_{\boldsymbol{\tau}}\right) \middle| \boldsymbol{\theta}(x)\right].$$

We note that the expectation in the second term can be factored into a product of expectations, none of which depend on $\boldsymbol{\tau}$ or $\boldsymbol{S}_{\boldsymbol{\tau}}$. The variance of this product is thus zero. We rewrite the remaining line using the identify $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$:

$$\mathrm{Var}\left(q_\star(x) \mid q\right) = q_0(x)^2 \times$$

$$\mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\mathbb{E}\left(\left(\prod_{A \ni x} ||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)}\right)^2 \middle| \boldsymbol{\theta}(x), \boldsymbol{\tau}, \mathbf{S}_{\boldsymbol{\tau}}\right) -\right.$$

$$\left.\left(\mathbb{E}\left(\prod_{A \ni x} ||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)} \middle| \boldsymbol{\theta}(x), \boldsymbol{\tau}, \mathbf{S}_{\boldsymbol{\tau}}\right)\right)^2 \middle| \boldsymbol{\theta}(x)\right].$$

Conditional on $\boldsymbol{\theta}(x)$ and $\boldsymbol{S}_{\boldsymbol{\tau}}$, the terms in both products are mutually independent, both *a priori* and *a posteriori*. Thus, we can factor the expectations:

$$\mathrm{Var}\left(q_\star(x) \mid q\right) = q_0(x)^2 \times$$

$$\mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\prod_{A \ni x} \mathbb{E}\left(\left(||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)}\right)^2 \middle| \theta(A), \tau(A), S_\tau(A)\right) -\right.$$

$$\left.\prod_{A \ni x}\left(\mathbb{E}\left(||A|| \left(\frac{\theta_\star(A)}{||A_\ell||}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{1-\theta_\star(A)}{||A_r||}\right)^{\mathbf{1}(x \in A_r)} \middle| \theta(A), \tau(A), S_\tau(A)\right)\right)^2 \middle| \boldsymbol{\theta}(x)\right].$$

We can now plug in our expressions for the expectations, calculated above:

$$\begin{aligned}
&\mathrm{Var}\left(q_\star(x) \mid q\right) = q_0(x)^2 \times \\
&\quad \mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau}}\left[\prod_{A \ni x} ||A||^2 \left(\frac{\theta(A)(\theta(A)\tau(A)+1)}{||A_\ell||^2(\tau(A)+1)}\right)^{\mathbf{1}(x \in A_\ell)}\right. \\
&\qquad\qquad \left(\frac{(1-\theta(A))((1-\theta(A))\tau(A)+1)}{||A_r||^2(\tau(A)+1)}\right)^{\mathbf{1}(x \in A_r)} - \\
&\qquad \left.\prod_{A \ni x} ||A||^2 \left(\frac{\theta(A)^2}{||A_\ell||^2}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{(1-\theta(A))^2}{||A_r||^2}\right)^{\mathbf{1}(x \in A_r)} \middle| \boldsymbol{\theta}(x)\right].
\end{aligned}$$

Since q is itself unknown, we take the expectation with respect to it:

$$\begin{aligned}
&\mathbb{E}_q\left[\mathrm{Var}\left(q_\star(x) \mid q\right)\right] = q_0(x)^2 \times \\
&\quad \mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}},\boldsymbol{\tau},\boldsymbol{\theta}(x)}\left[\prod_{A \ni x} ||A||^2 \left(\frac{\theta(A)(\theta(A)\tau(A)+1)}{||A_\ell||^2(\tau(A)+1)}\right)^{\mathbf{1}(x \in A_\ell)}\right. \\
&\qquad\qquad \left(\frac{(1-\theta(A))((1-\theta(A))\tau(A)+1)}{||A_r||^2(\tau(A)+1)}\right)^{\mathbf{1}(x \in A_r)} - \\
&\qquad \left.\prod_{A \ni x} ||A||^2 \left(\frac{\theta(A)^2}{||A_\ell||^2}\right)^{\mathbf{1}(x \in A_\ell)} \left(\frac{(1-\theta(A))^2}{||A_r||^2}\right)^{\mathbf{1}(x \in A_r)}\right].
\end{aligned}$$

We call this expectation the variance function. Finally, we note that conditional on $\mathbf{S}_{\boldsymbol{\tau}}$, $\theta(A), \tau(A)$ are independent of $\theta(A'), \tau(A')$ for any two distinct regions A and A'.

This allows us to rearrange our expectation to obtain

$$\mathbb{E}_q\left[\mathrm{Var}\left(q_\star(x) \mid q\right)\right] = q_0(x)^2 \times$$

$$\mathbb{E}_{\mathbf{S}_{\boldsymbol{\tau}}}\left[\prod_{A \ni x} ||A||^2 \mathbb{E}_{\theta(A),\tau(A)}\left(\frac{\theta(A)(\theta(A)\tau(A)+1)}{||A_\ell||^2(\tau(A)+1)}\middle| S_\tau(A)\right)^{\mathbf{1}(x\in A_\ell)}\right.$$

$$\left.\mathbb{E}_{\theta(A),\tau(A)}\left(\frac{(1-\theta(A))((1-\theta(A))\tau(A)+1)}{||A_r||^2(\tau(A)+1)}\middle| S_\tau(A)\right)^{\mathbf{1}(x\in A_r)} - \right.$$

$$\prod_{A \ni x} ||A||^2 \mathbb{E}_{\theta(A),\tau(A)}\left(\frac{\theta(A)^2}{||A_\ell||^2}\middle| S_\tau(A)\right)^{\mathbf{1}(x\in A_\ell)}$$

$$\left.\mathbb{E}_{\theta(A),\tau(A)}\left(\frac{(1-\theta(A))^2}{||A_r||^2}\middle| S_\tau(A)\right)^{\mathbf{1}(x\in A_r)}\right].$$

The expectations with respect to $\theta(A)$ and $\tau(A)$ can be estimated during the same quadrature routines used to compute the posterior distributions of $\theta(A)$, described in Section 2.5.2. The expectation with respect to $\boldsymbol{S_\tau}$ can then be calculated during the forward-backward routine used to calculate the posterior distribution of $\boldsymbol{S_\tau}$, as described in the same section.

We naturally expect more absolute variation between samples in areas where the densities of all the samples are higher, so we prefer a standardized dispersion function measuring the coefficient of variation. The posterior mean coefficient of variation of $q_\star$ at any given point is not analytically tractable; we can obtain an estimate by taking the square root of the variance function and dividing by the mean density function. We illustrate the application of the dispersion functions in Section 2.8.1.

2.7.2 Dirichlet Process Mixture of HAPT

In many applications we may not believe that the samples collected all share a single common structure. A more appropriate model may be that the samples are drawn from several latent populations, with samples being drawn from the same population having structure in common. In this case we may reconstruct the latent structure by

clustering the samples. To learn the clustering of samples without fixing the number of clusters in advance, we add a Dirichlet process component to the model. We can write the model as follows:

$$\begin{aligned} X_{ij} \mid Q_i &\overset{iid}{\sim} Q_i \\ Q_i \mid Q_i^*, \boldsymbol{\tau}_i^* &\overset{iid}{\sim} \mathrm{PT}(Q_i^*, \boldsymbol{\tau}_i^*) \\ (Q_i^*, \boldsymbol{\nu}_i^*, \boldsymbol{\tau}_i^*) \mid G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(\alpha H(Q^*, \boldsymbol{\nu}^*, \boldsymbol{\tau}^*)), \end{aligned}$$

where the base measure can be factored as

$$\begin{aligned} H(Q^*, \boldsymbol{\nu}^*, \boldsymbol{\tau}^*) &= [\pi(Q^* \mid \boldsymbol{\nu}^*) \times \pi(\boldsymbol{\nu}^*)] \times \pi(\boldsymbol{\tau}^*) \\ &= [\mathrm{PT}(Q_0, \boldsymbol{\nu}^*) \times \mathrm{SIS}(\boldsymbol{\Gamma}_{\boldsymbol{\nu}}^*)] \times \mathrm{SIS}(\boldsymbol{\Gamma}_{\boldsymbol{\tau}}^*). \end{aligned}$$

The Dirichlet process introduces clustering among the samples, so that some set of Q_i, belonging to the same cluster, share a corresponding Q_i^*, $\boldsymbol{\tau}_i^*$, and $\boldsymbol{\nu}_i^*$. Conditional on the clustering structure, the model reduces to a collection of independent HAPT models.

We call this model a Dirichlet Process Mixture of Hierarchical Adaptive Polya Trees, or DPM-HAPT. It is comparable to the Nested Dirichlet process in the way it induces clustering among the samples, but is considerably more flexible. While the NDP requires that all samples in a cluster have identical distributions (MacEachern, 2008), DPM-HAPT allows the distributions within cluster to vary according to the HAPT model. In addition, the advantages of the HAPT model discussed earlier, such as flexible modeling of variation in different parts of the sample space, still apply.

DPM-HAPT requires one adjustment to the prior specification of $\boldsymbol{\tau}$. While Ma (2017) recommends a range of $(-1, 4)$ (on the $\log_{10}$ scale) for the precision parameter, very small values of $\tau(A)$ do not make sense in the context of clustering, as they

imply bimodal clusters. To avoid bimodality it is desirable to have $\theta(A)\tau(A) > 1$ for reasonable values of $\theta(A)$; we recommend restricting the range of $\log_{10}\tau(A)$ to the interval $(1, 4)$, which satisfies that requirement as long as the true distributions are not too far from the prior common measure.

Posterior computation for HAPT-DPM consists of a combination of standard Dirichlet Process methods and the HAPT posterior calculations described in Section 2.5.1. As noted above, conditional on the clustering structure, the model consists of a number of independent HAPT models. Although the HAPT model is not fully conjugate, our posterior computation strategy allows us to calculate the marginal likelihood with arbitrary precision. This allows the use of a Dirichlet process algorithm designed for conjugate mixture models. We use the Chinese Restaurant Process representation to sample the clustering structure, using marginal likelihoods calculated from the HAPT model.

This algorithm requires the computation of

1. marginal likelihoods under the HAPT model for clusters including a single sample, and

2. marginal posterior predictive likelihoods under the HAPT model of one sample conditional on one or more other samples making up a cluster.

The first item is straightforward. The second is easily achieved by fitting the HAPT model twice, and calculating

$$f(\boldsymbol{X}_i \mid \boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_k}) = \frac{f(\boldsymbol{X}_i, \boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_k})}{f(\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_k})}.$$

2.8 Simulation results

2.8.1 Estimation of the dispersion function

In order to obtain data with dispersion that varies across the sample space, we simulate from a mixture of three Beta distributions, such that the variation between sample densities is low in the middle of the space and much higher near zero and one. Each sample is a mixture of three components: Beta(2,2), Beta(1,12), and Beta(12,1). The corresponding weights w_1, w_2, w_3 of the three components are drawn according to the following scheme. First we draw $w_1 \sim \text{Beta}(80, 20)$. Then we draw $v \sim \text{Beta}(1, 1)$ and set $w_2 = v(1 - w_1)$ and $w_3 = (1 - v)(1 - w_1)$. This results in the central part of the sample space having a small amount of variation between samples, while the edges on either side have much more variation.

One hundred sample densities are plotted in Figure 2.4(a). The variation in the dispersion of the sample densities can be seen clearly.

Figure 2.4(b) shows the estimated centroid and sample densities from fitting the HAPT model. The dispersion function, or estimated coefficient of variation is plotted in Figure 2.4(c). The dispersion function clearly shows how the variation between samples is low near the center of the space and high on either end.

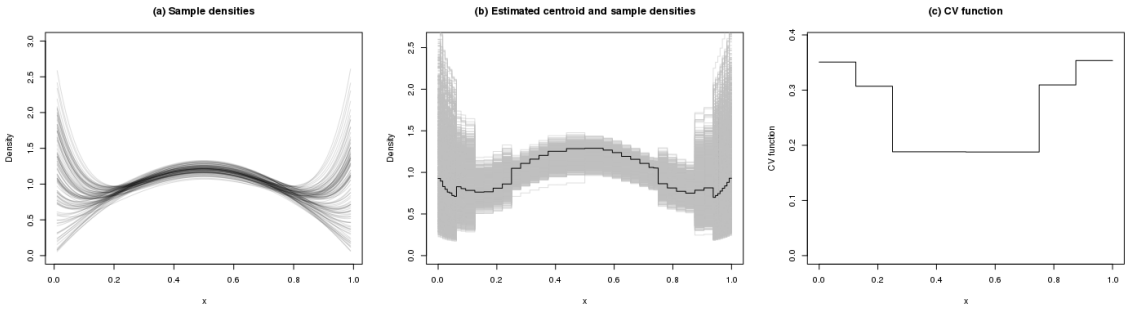

FIGURE 2.4: (a) One hundred sample densities from the simulation setting used in Section 2.8.1; (b) Estimated centroid and sample densities after fitting the HAPT model; (c) Estimated coefficient of variation function.

2.8.2 DPM-HAPT Simulations

We simulate a simple 1-dimensional example to demonstrate the clustering behavior of the DPM-HAPT model. The simulation contains 30 samples belonging to three true clusters, with 15, 10, and 5 samples respectively. Each sample is drawn from a mixture of a Uniform(0,1) distribution and a Beta distribution, with the parameters of the Beta varying by cluster: Beta(1,5) for the first cluster, Beta(3,3) for the second cluster, and Beta(5,1) for the third cluster. The weights of the two components are randomized in each sample. The weight of one component is drawn from a Beta(10,10) distribution, which creates weights varying approximately between 0.3 and 0.7, with the actual observed proportions in realized samples varying more widely due to the additional Binomial variation. Sample densities are plotted in Figure 2.5(a). Each sample contains $n = 300$ points.

An MCMC sampler is run using the Chinese Restaurant Process to sample the clustering structure. We summarize the results by looking at how often each of the 30 samples is clustered together with each other sample. These results are plotted in Figure 2.5(b). We can see in the figure that the DPM-HAPT model clearly identified the three clusters.

We now consider an example in which the variation between samples differs across the sample space. Samples with heterogeneous variation are drawn from mixtures of four Beta distributions, with varying parameters for each cluster. The parameter are chosen so that each cluster has much higher variation between samples on the right half of the space than on the left half. Each sample is a mixture of four components, shown in Figure 2.6: Beta(1,6), Beta(2,5), Beta(5,2), and Beta(6,1). The corresponding weights w_1, w_2, w_3, and w_4 of the mixture components are drawn according to the following scheme. Let v_1, v_2, v_3 be Beta random variables whose parameters that differ by cluster, with v_3 having a much larger variance than v_1 and

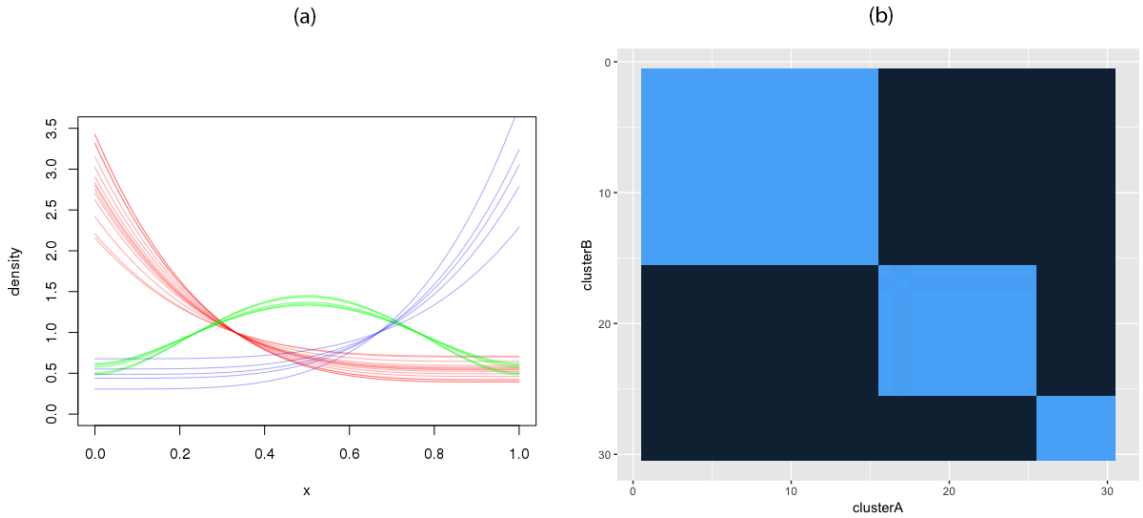

FIGURE 2.5: (a) Sample densities for the clustering simulation, and (b) probability of two samples clustering together based on the DPM-HAPT model, which clearly identifies the three clusters that exist in this simulation, despite significant variation between samples within clusters.

v_2. Then we calculate the weights as follows:

$$w_1 = v_1 v_2$$

$$w_2 = v_1(1 - v_2)$$

$$w_3 = (1 - v_1)v_3$$

$$w_4 = (1 - v_1)(1 - v_3).$$

Because v_3 has a much larger variance than v_1 and v_2, we end up with much more variation between densities on the right half of the space than on the left half.

We consider 30 samples belonging to three true clusters, as above. Figure 2.7 (a) shows one hundred draws from each of three clusters used in this simulation, to illustrate the variability between and within clusters and the heterogeneity across the sample space. We can see that the clusters have substantially more within-cluster variation on the right half of the space than on the left half.

As in the previous example, we simulate three clusters with 15, 10 and 5 groups respectively. We draw 150 observations from each group, and apply DPM-HAPT to cluster the resulting samples. We run the MCMC for 1,000 draws after burnin; estimated coclustering probabilities are plotted in Figure 2.7 (b). The three true clusters are clearly identified even in the presence of substantial heterogeneity.

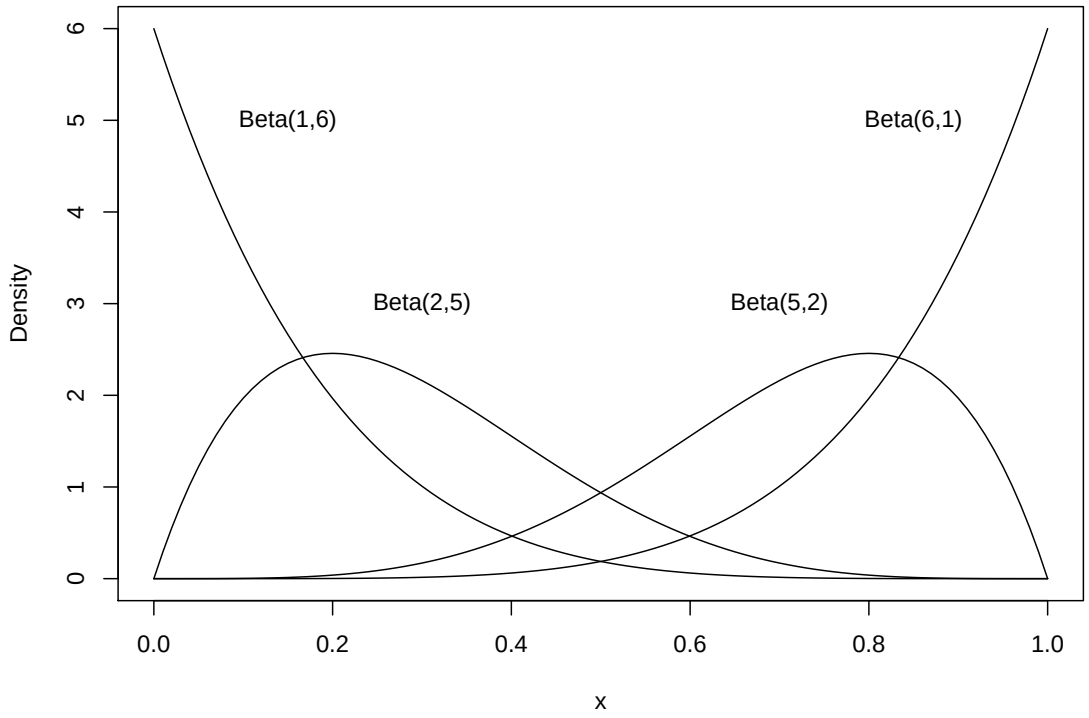

FIGURE 2.6: The four mixture components used in the simulation in Sections 2.8.1.

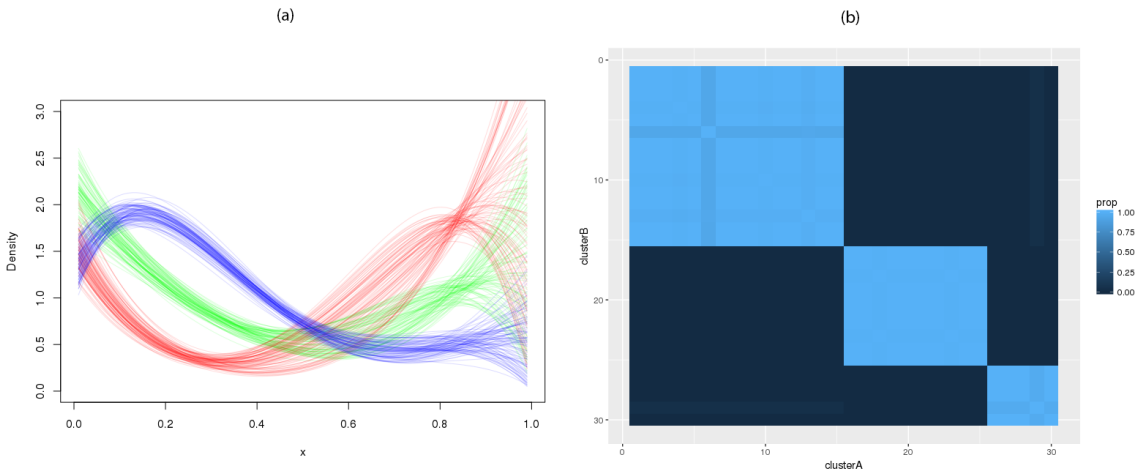

FIGURE 2.7: (a) One hundred draws from each of three clusters in the heterogeneous variation clustering example; (b) Probability of two samples clustering together based on 1,000 MCMC draws for the heterogeneous variance example. Despite the variation in the dispersion of the densities, HAPT-DPM clearly identifies the three true clusters.

2.9 Application: DNase-seq profile clustering

DNase sequencing (DNAse-seq) is a method used to identify regulatory regions of the genome (Song and Crawford, 2010). DNA is treated with Deoxyribonuclease (DNase) I, an enzyme that cuts the DNA strand. The cut strands are then sequenced and the locations of the cuts are identified and tallied. The vulnerability of the DNA strand to DNase varies by location, resulting in a distribution of cut counts which is nonuniform. The density of this distribution is related to various biological factors of interest; for example, it tends to be high near potential binding sites for transcription factors, since these proteins require access to the DNA strand in much the same way as DNase I, but will be low if a transcription factor already bound at that site blocks access to the DNA strand.

We consider the problem of clustering DNase-seq profiles near potential transcription binding sites, identified by a specific genetic motif. Each sample consists of observed counts in a range of 100 base pairs on either side of one occurrence of the motif. A single motif, consisting of 10–20 base pairs, may appear tens of thousands of time in the genome, with each occurrence presenting one sample for analysis. Many samples, however, have very few cuts observed. For analysis we restrict ourselves to samples which meet a minimum sample size threshold.

Different locations where the transcription factor motif of interest appears may be expected to show different DNase behavior in the region around the motif for a variety of reasons. This makes clustering a more appropriate approach to the problem than treating all the samples as having a single common structure. Identifying clusters of locations which have similar DNase-seq profiles may reveal previously unrecognized factors. We also expect within-cluster variation beyond simple sampling variation, which makes the Nested Dirichlet process unsuitable.

Here we present data from the ENCODE project (ENCODE Project Consortium,

2012) for locations surrounding a motif associated with the RE1-silencing transcription factor (REST). REST suppresses neuronal genes in non-neuronal cells (Chong et al., 1995). The data includes 48,549 locations where the REST motif appears in the genome. The motif consists of 21 base pairs, and the data includes an additional 100 base pairs on each side, for a total of 221 base pairs. In all, 922,704 cuts were tallied, an average of 19 per location. 468 locations have zero cuts observed. The number of cuts per location is heavily right skewed, with a median of 13 observations, first and third quartiles of 7 and 21 respectively, and a maximum of 2,099 cuts observed in a single sample.

For this analysis we restrict ourselves to locations which have at least 200 observations, a total of 265 samples. These samples include a total of 70,225 observations, an average of 330 observations per sample. The distribution is still quite skewed, with a minimum of 201 observations and a maximum of 2099. The median is 279 and the first and third quartiles are 232 and 366 observations.

We fit the DPM-HAPT clustering model to this data, using 100 post-burnin draws for inference. The model estimates 7 clusters with high probability (see Figure 2.8(a)), of which there are three large clusters and four singleton locations, each of which consists of a single large spike. One of the singleton locations occasionally joins one of the larger clusters, resulting in 6 clusters. A heatmap of the clustering structure is shown in Figure 2.8(b). The clusters are clearly differentiated and vary in size, with the largest cluster containing about 130 locations, though the cluster sizes vary from iteration to iteration due to uncertainty in the cluster assignment.

The estimated mean densities of the three largest clusters are plotted in Figure 2.9. One of these clusters includes locations with cuts which are roughly symmetric around the transcription factor binding site, while the other two largest clusters include locations which have cuts heaped up on one side or another of the binding site. Additional clusters show other densities which do not conform to the general

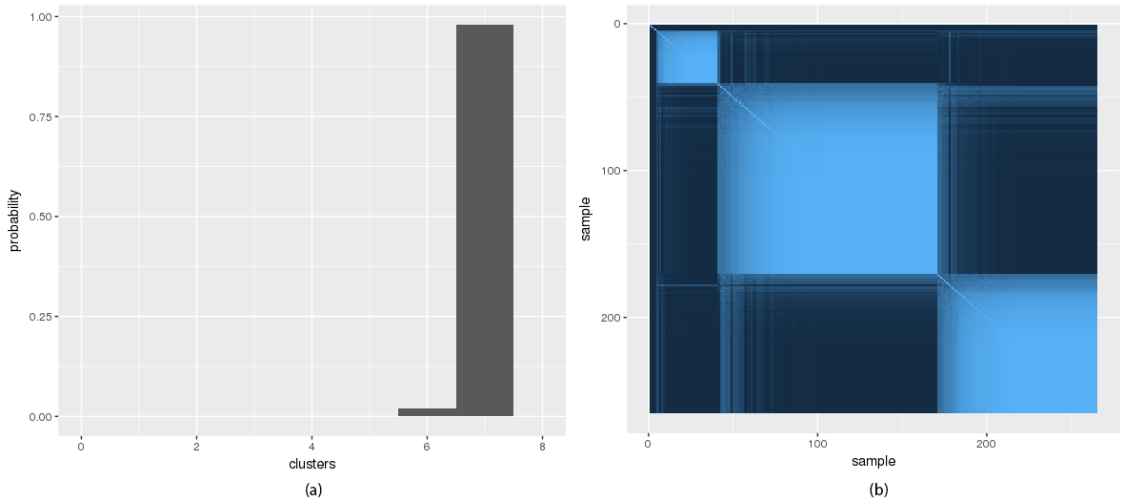

FIGURE 2.8: (a) Estimated number of clusters from the DPM-HAPT in the DNase-seq application; (b) the model shows clear clustering of the samples.

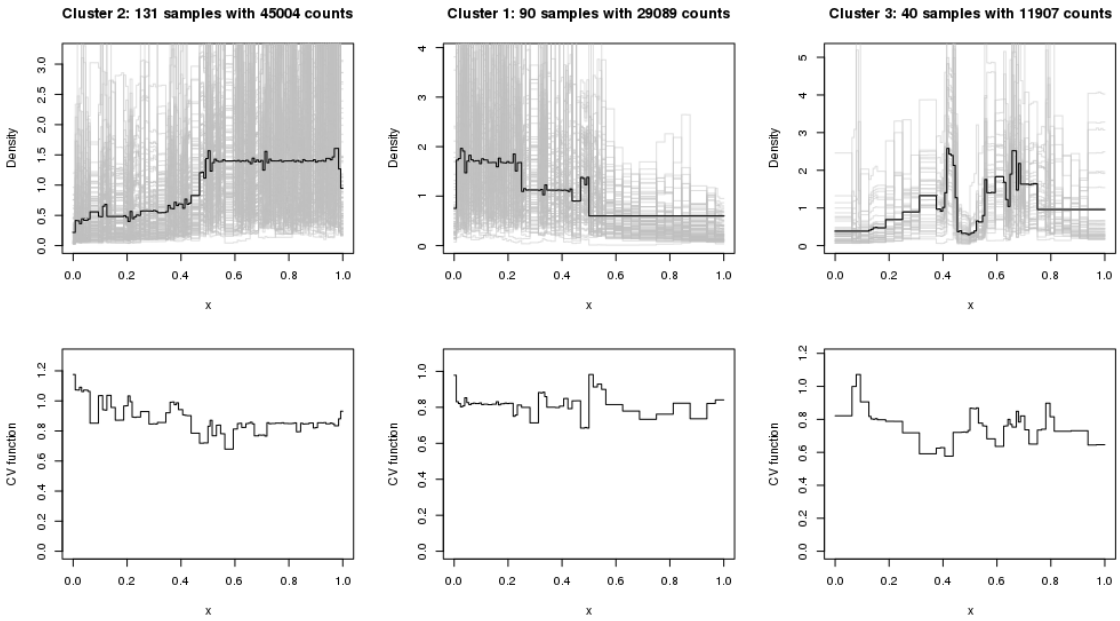

FIGURE 2.9: Posterior mean densities and estimated dispersion functions of the three largest clusters in the DNase-seq example. The heavy black line shows the cluster centroid; light gray lines in the background show the estimated means of each sample in the cluster. The dispersion function is plotted below.

patterns of the three largest clusters. The plots show substantial variation around the cluster centroid, much more than can be explained by sampling variation alone. The estimated dispersion functions are quite noisy; this is not surprising given that they are between-sample dispersions estimated with only 40–50 degrees of freedom.

2.10 Discussion

The HAPT model offers a compelling alternative to existing nonparametric models that share information across multiple samples. The Polya tree's ability to directly model the density of an absolutely continuous distribution frees us from the necessity of using mixture models to obtain densities, while the computational tractability of the posterior avoids the need to run MCMC chains for posterior inference. The addition of the SIS prior allows us not only to more accurately model the densities of interest, but also to estimate a fully nonparametric dispersion function over the sample space. The model extends easily—both conceptually and computationally—to the setting where we do not believe all our samples have the same common structure, where DPM-HAPT allows us to learn both clustering structure and the distributional structure within each cluster.

Although we have presented HAPT in a one-dimensional space for the sake of clarity, adoption of the randomized recursive partitioning scheme first introduced in Wong and Ma (2010) allows the extension of HAPT to model densities in multidimensional spaces. Variables other than simple continuous ones can also be handled naturally—all that is needed is the definition of an appropriate recursive partition. This allows inclusion of categorical and ordinal-valued variables, as well has more exotic possibilities: a continuous variable that lives on the surface of a torus, a partially ordered categorical variable, or a zero-inflated variable with a point mass at zero and a continuous component on the positive halfline.

The Polya tree's decomposition of the density space into orthogonal Beta-distributed

random variables, which extends to HAPT, is central to HAPT's computational efficiency. It also allows the performance of quick online updates in the HAPT model: when a new data point arrives we only need to update the nodes of the Polya tree which contain the new data point, rather than recomputing the entire posterior. In a HAPT truncated at a depth of L levels, this means we need to reevaluate the posteriors of only L nodes, rather than 2^L. HAPT may thus be used in streaming data settings where fast online updates are essential.

The DPM-HAPT model may also be easily extended by replacing the Dirichlet process by any clustering-inducing process which admits inference given the marginal likelihoods and conditional probabilities of the clusters. This includes, for example, the Pitman-Yor process. This allows the properties of the clustering process to be adapted if the clustering assumptions implicit in the Dirichlet process are not appropriate.

3

On the use of marginal likelihoods

In posterior computations for stateful Polya trees, marginal likelihoods under each state are combined with prior probabilities to obtain the posterior probability of each state. This is analogous to the use of marginal likelihoods for Bayesian model comparison, selection, and hypothesis testing (often through derivative quantities such as posterior probabilities or Bayes factors). Although a number of alternative approaches to Bayesian model comparison and hypothesis testing have been advanced over the years, recent examples including Bernardo et al. (1999); Kamary et al. (2014), methods using quantities derived from the marginal likelihood continue to dominate the literature.

Marginal likelihoods evaluate the strength of a model based on how well the prior predictive distribution fits the observed data, and are sensitive to the prior distribution used within a given model structure (de Santis and Spezzaferri, 1999; Conigliani and O'Hagan, 2000). Berger and Guglielmi (2001) recognize this sensitivity in conducting a two-sample test with a nonparametric (Polya tree) alternative, and calculate the Bayes factor over a range of values for a multiplicative factor on the Polya tree concentration parameter (see Figure 3.1). Over the range considered, the

resulting Bayes factor varies by over three orders of magnitude, casting considerable doubt on what conclusions that can be drawn from a single analysis with a fixed prior.

In addition to being sensitive to nuisance parameters in the prior specification, I argue that marginal likelihoods and quantities derived from them are conceptually the wrong way to evaluate models. I expand on these points below, and offer a suggestion of how to better compare Bayesian models and conduct Bayesian hypothesis tests. In Section 3.1 I argue that using marginal likelihoods is fundamentally flawed, as they evaluate a model which is not actually the target of our inference, and is often not a good proxy for the target of inference. This often results in models which can fit the data well (as measured by the likelihood) being excessively penalized for the *prior* not fitting the data well, a phenomenon known as Lindley's paradox. In Section 3.2 I discuss one approach to resolving the problems with marginal likelihoods, based on splitting the data (or the likelihood) into training and test samples. This approach has previously appeared in the literature in the context of evaluating models with improper priors, but not applied to resolving Lindley's paradox. In Section 3.3 I show results from simulations demonstrating that the training/test split does not hurt the discriminatory power of the Bayesian hypothesis test. In Section 3.4 I show how the training/test split can be applied to a stateful Polya tree to improve the accuracy of posterior inference, and Section 3.5 contains some general discussion.

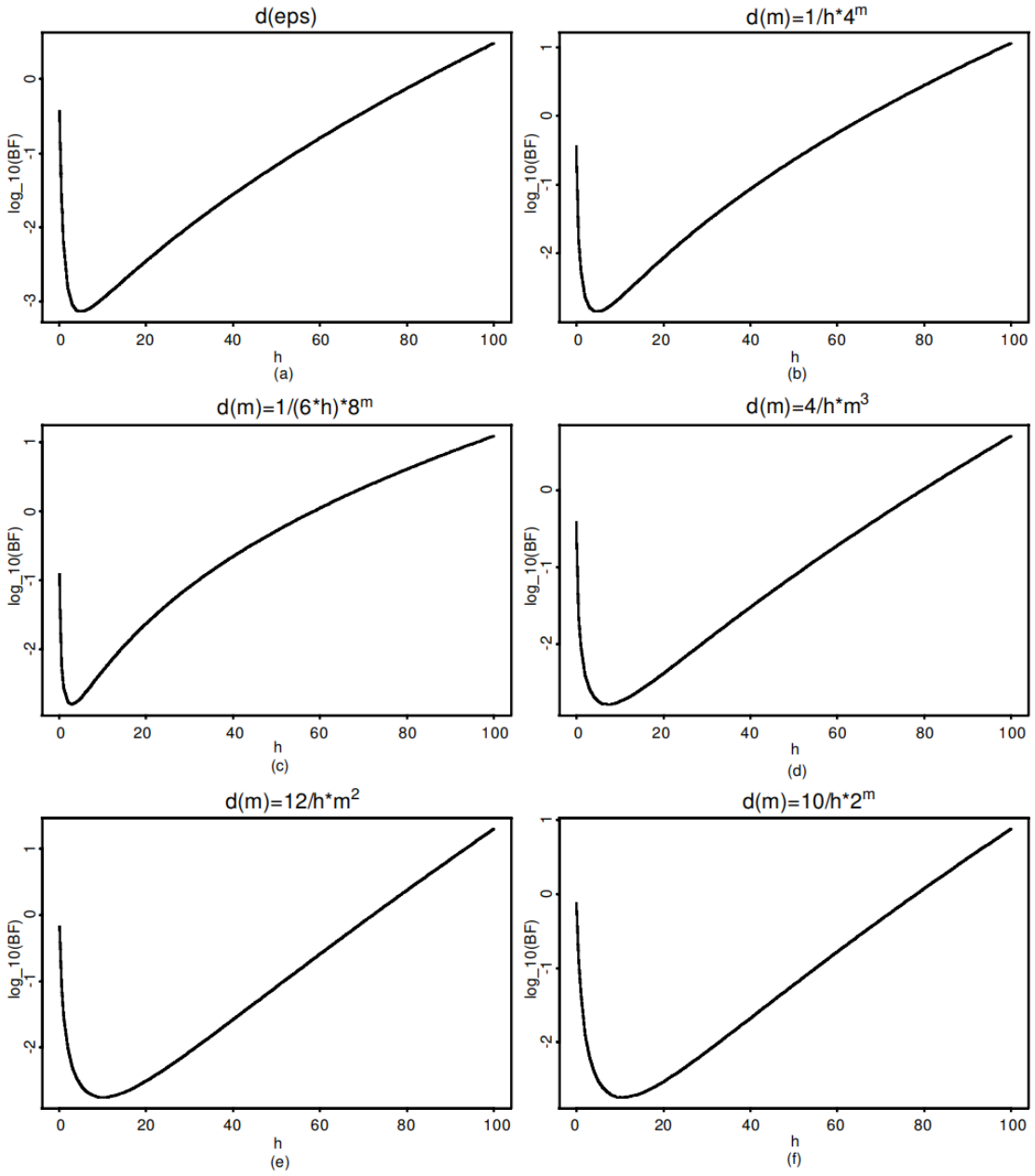

FIGURE 3.1: Figure 1 from Berger and Guglielmi (2001), evaluating the sensitivity of the Bayes factor to the choice of the multiplicative factor on the concentration parameter. Each pane corresponds to a different choice for the growth rate of the concentration parameter as one moves down the tree; for a given growth rate, a range of values for the multiplicative factor h are considered. In all cases it can be seen that the conclusion is strongly dependent on the multiplicative factor used—note that the vertical axis is on a $\log_{10}$ scale.

3.1 The marginal likelihood aims at the wrong target

The marginal likelihood evaluates the fit of a fully-specified joint model for data x and parameters θ. In practice, whether doing hypothesis testing or model selection, this joint model is rarely the object that we are actually interested in doing inference on. The target for inference is almost always

1. a hypothesis about parameter values, without a fully specified probabilistic model; or

2. a conditional probabilistic model for the data given parameters.

The following example, following an early Bayesian analysis by Laplace, illustrates the point. For the years 2013-2015, the three most recent years of data available at the time of writing, the Center for Disease Control and Prevention recorded 12,926 births registered in Durham county, of which 6,604 were male (CDC, 2017). Suppose, in ignorance of hundreds of years of data on human birth sex ratios, we wanted to test whether or not the proportion of male births in Durham was equal to $\frac{1}{2}$. That is, we wish to test the hypothesis

$$H_0 : \pi = 0.5 \qquad H_1 : \pi \neq 0.5,$$

.

A frequentist hypothesis test gives a two-sided p-value of 0.013, which would lead most practitioners to reject the null hypothesis. Laplace's approach evaluates the posterior distribution for the proportion of male births π under a uniform prior, obtaining a Beta(6,605, 6,323) posterior distribution. This gives

$$P(\pi > 0.5 \mid X) = .993,$$

which would have lead Laplace to conclude that the sex ratio at birth is almost

certainly greater than 0.5, with the posterior distribution centered at approximately 0.51.

The modern Bayesian approach posits two models

$$M_0 : \qquad X \sim \text{Binomial}(0.5)$$

$$M_1 : \qquad X \mid \pi \sim \text{Binomial}(n, \pi)$$
$$\pi \sim \text{Beta}(1, 1)$$

and calculates the Bayes factor in favor of the null, which is 4.19. If we had assigned equal prior probabilities to the two models, we would find that the posterior probability of the null model had climbed to 0.81, while the probability of the alternative model has fallen to 0.19.

Such discrepancies between Bayesian and frequentist analyses are familiar in the literature, and are attached to various combinations of the names Jeffreys, Lindley, and Bartlett. Lindley (1957) demonstrated that there may be an arbitrarily large discrepancy between frequentist and Bayesian conclusions based on a data set, in the sense that for every arbitrarily small pair $\epsilon, \delta > 0$, it is possible to construct an example such that the frequentist will reject the null hypothesis at level ϵ while the Bayesian will assign it a posterior probability of at least $1 - \delta$. Such examples are not merely academic; they can be characterized as arising when the effect is small and the sample size is large, a regime of increasing interest in an era of "big data." As demonstrated in the example above, the sample size need not be unrealistically large, nor the effect excessively small, for the discrepancy to be observed. The frequentist approach rejects the null because the maximal likelihood under the alternative provides a substantially better explanation for the observed data, while the Bayesian hypothesis testing approach awards the null most of the posterior probability because *most* (under the prior measure) of the points in the alternative parameter space ex-

plain the data much worse than the null hypothesis does. The two approaches are comparing different objects: the frequentist approach compares conditional models, while the Bayesian approach compares joint models.

I argue that the *joint model is the wrong inferential target.* While we often blithely assert an equivalence between the models M_0, M_1 and the hypotheses H_0, H_1, we did not set out to test whether the birth rate was 0.5 or Uniform(0,1), as the Bayesian hypothesis test does. The specification of a prior is a nuisance parameter which we would gladly be rid of—particularly in high-dimensional settings such as stateful Polya trees. It is precisely this desire to be rid of the nuisance that has led to an active field of research into "objective priors" for Bayesian hypothesis testing; see e.g. Bernardo and Rueda (2002); Pericchi (2005); Bayarri and García-Donato (2008).

Inference on joint models is useful only to the extent that the joint models are accurate representations of our beliefs about the hypotheses or conditional models which are of interest. Proposing various "objective" priors does not help if none of the resulting joint models are good proxies for our believes about the conditional models. All these joint models are subject to the prior predictive putting most of the mass in areas not supported by the data, resulting in low marginal likelihoods through no fault of the conditional model or hypothesis which is actually our inferential target.

3.2 Avoiding penalization due to prior uncertainty

To resolve the issue of prior parameter uncertainty penalizing the conditional model, we set aside a portion of the data as a training set. We update the prior using this training data to obtain a *partial posterior* distribution; we then use that partial posterior as the model to evaluate the fit of the remainder of the data, which we refer to as the test set. This approach is not new, and when used to calculate Bayes factors has been referred to in the literature as the *partial Bayes factor* (see O'Hagan (1995, 1997); De Santis and Spezzaferri (1997)). A special case is the intrinsic Bayes

factor of Berger and Pericchi (1996). O'Hagan (1995) introduces the fractional Bayes factor, which is an idealization of the partial Bayes factor: rather than splitting the data into test and training, we factor the likelihood into train an test parts:

$$\mathcal{L}(\theta; X) = \mathcal{L}(\theta; X)^{n_{\text{train}}/n} \times \mathcal{L}(\theta; X)^{n_{\text{test}}/n}.$$

This avoids the issue of variation in the train/test data split.

There are a number of considerations when determining the size of the training and test sets. We consider an example discussed by Jefferys (1990). The data consist of 104,490,000 Bernoulli trials, of which 52,263,471 were "successes." We wish to test the null hypothesis that $\theta = 0.5$ against the alternative $\theta \neq 0.5$. The standard frequentist test rejects the null with a two-sided p-value of 0.0003, while the Bayesian approach, with a uniform prior on θ under the alternative, gives a Bayes factor of approximately 11.9 in favor of the null. That is, if we assume that each of the two hypothesis is equally likely *a priori*, the Bayesian analysis concludes that the posterior probability of the null is in excess of 0.92. Due to the very large sample size, the likelihood in this case is very peaked around the maximum likelihood estimate $\hat{\theta} \approx 0.500177$. Although values of θ very near this estimate provide substantially better likelihoods than the null, the vast majority of the prior weight under the alternative—approximately 99.96%—is placed on values of θ which explain the data worse than the null model.

Figure 3.2 shows the log Bayes factors in favor of the null calculated from 1000 simulations of training and test sets as the training sample sizes varies from 0 to $n = 104,490,000$. Individual simulation traces are displayed faintly in grey. The mean (on the log Bayes factor scale) of the 1000 simulations is plotted as the upper red curve. All the simulations begin on the left axis, where the training size is zero, with a log Bayes factor of 2.48, corresponding to the 11.9 odds given earlier. As the training set increases and the prior concentrates, they dip sharply below zero before

leveling off. Eventually every trace returns to 0 when the entire data set has been set aside for training and no data is available for testing, as we are then unable to make any distinction between models. The sharp dip is characteristic of scenarios where Lindley's paradox appears, and demonstrates that the standard Bayes factor's apparent preference for the null is in fact merely an artifact of the prior distribution under the alternative.

Several additional curves are plotted. The upper blue curve gives the standard deviation of the 1000 simulated traces at each point. The fairly large standard deviation indicates that we cannot rely on a single training/test split; probably we will want to split several times and average the results. The black curve gives the fractional Bayes factor, which follows the mean of the simulated traces closely but tends to be slightly more in favor of the alternative.

Finally, three more curves show how the fractional Bayes factor can be approximately decomposed into two regimes. We may imagine that rather than removing one point from the test set and placing it in the training set, we have a separate test set of fixed size—identical to the observed data—and we slowly increase the training set without touching the test set. This gives the monotonically decreasing red curve. If we were to continue increasing the size of the training set, asymptotically this curve would approach the lower grey line, which is the log likelihood ratio between the null and the MLE under the alternative.

Conversely, we may imagine having a fixed training set identical to the observed data, and increasing the size of the test set. This gives (taking into account that the further we are to the right on the plot axes, the smaller the test set) the monotonically increasing red line. It is well known that under the alternative hypothesis, the log Bayes factor asymptotically grows linearly with sample size, so it is not surprising that this second curve is very close to a straight line.

Finally, the blue curve following the black fractional likelihood curve is an ap-

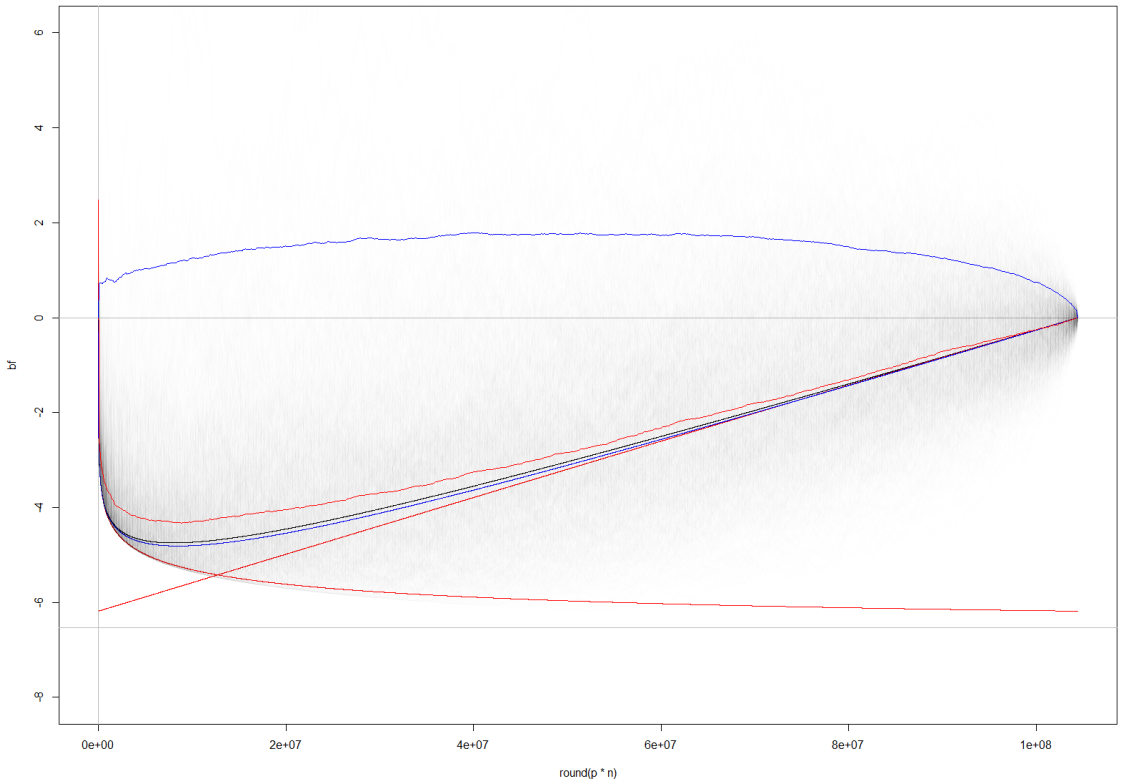

FIGURE 3.2: Simulation of partial and fractional Bayes factors for different training set sizes.

proximation of the fractional likelihood based on a combination of these two red curves. While the red curves represent hypothetical situations, not corresponding to reality, in which one of the training/test sets is increasing while the other is fixed, they are able to approximate the fractional Bayes factor quite well. For training sets including a small fraction of the data, removing a point from the very large test set is a miniscule change, so the behavior of the fractional Bayes factor curve is in a regime where its behavior is dominated by the effect of the training set, and it follows the decreasing red curve closely. Conversely, for large training sets, adding a point the the training set is a miniscule change, so the fractional Bayes factor curve's behavior is dominated by the effect of the test set, which follows the increasing red curve. The approximation is worst for moderate training set sizes.

This decomposition into two regimes should inform our choice of the size of the training set. We want to provide enough training data to concentrate the prior to a reasonable degree, without falling into the second regime where decreasing the test set size leads to linearly decreasing ability to discriminate between models.

3.3 Simulations

There may be some concern that by giving up part of the data as a training set, the test between null and alternative hypotheses will lose power. Simulations show this not to be the case. Training sets can be quite small relative to the entire data set, particularly when samples sizes are large, and the loss of test set data is compensated by the increased precision of the models being tested.

In Figure 3.3 we plot received operating characteristic curves for three Bayesian tests, the classical likelihood ratio test, and two reference curves. The Fractional and Partial Bayes factor tests perform identically to the standard Bayes Factor and the classical likelihood ratio test. The Partial Bayes Factor test is based on the average log partial Bayes factor over 100 random splits of training and test data. A Bayes

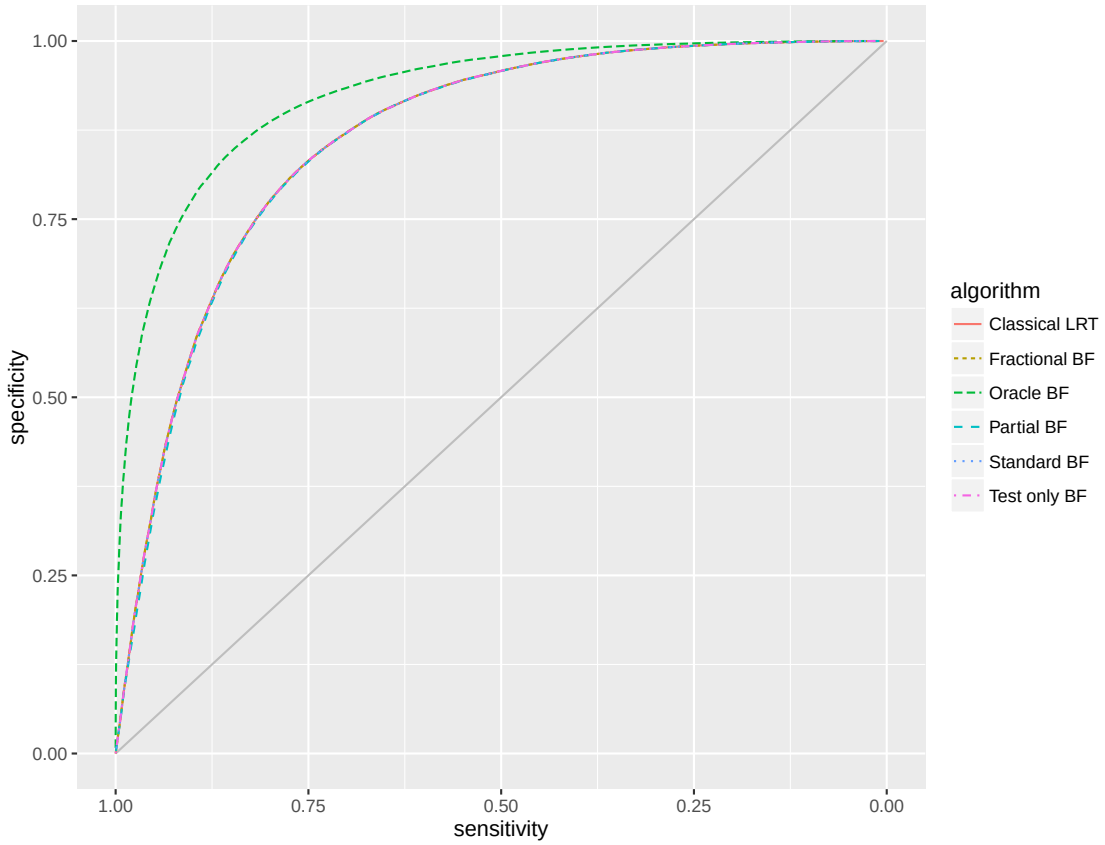

FIGURE 3.3: Received operating characteristics for various tests in the simulation setting discussed in 3.3. The sample size is 104,490,000, with the training sample size set according to the square root (approximately 10000). The true effect is 0.5001, tested against a null of 0.5. 100000 simulation runs were used to compute the curves.

factor calculated only on the test set, ignoring the training set entirely, is marginally worse, though the difference cannot be seen on this plot due to the small size of the training set. Figure 3.4 shows the results of the same simulation setting but with an artificially large training sample of 10,000,000. Here we see the loss of power in the test-only Bayes factor, while the partial and fractional Bayes factors retain the same power as the traditional Bayes factor.

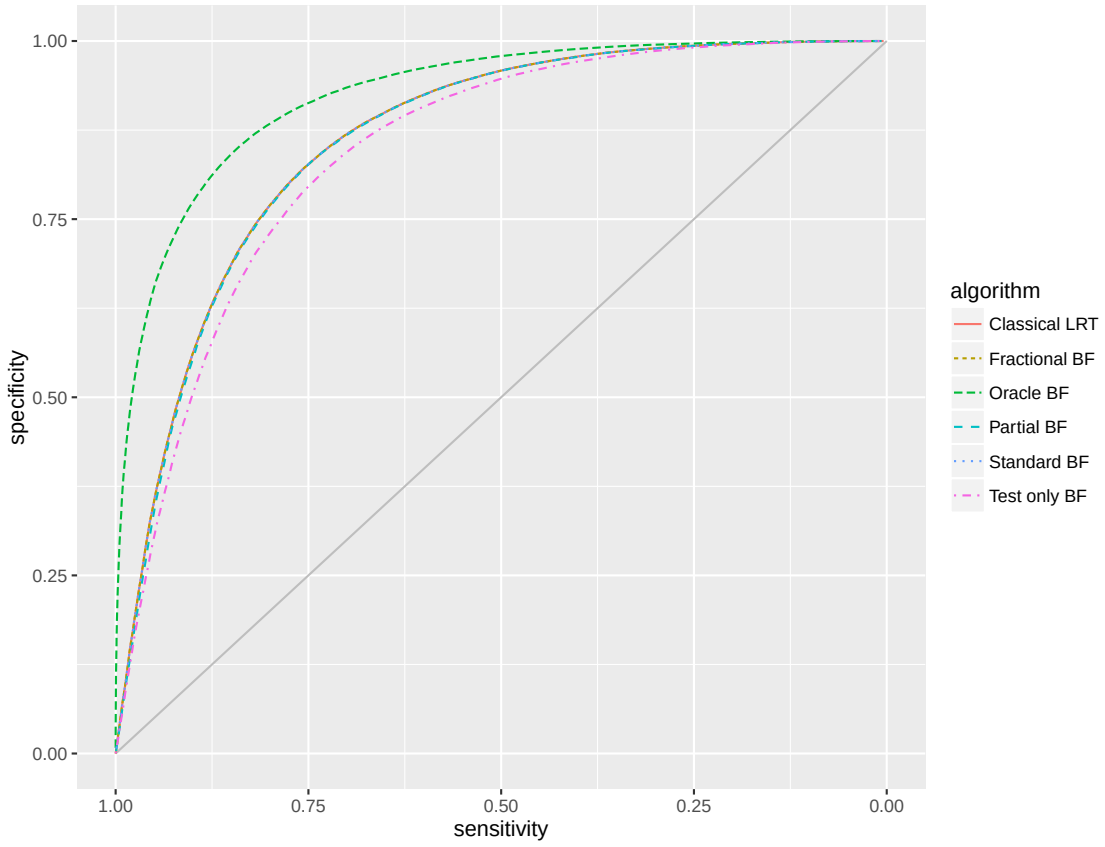

FIGURE 3.4: Received operating characteristics for various tests in the simulation setting discussed in 3.3. The sample size is 104,490,000, with an artificially large training set of 10,000,000. The true effect is 0.5001, tested against a null of 0.5. 100000 simulation runs were used to compute the curves.

3.4 Application to stateful Polya trees

Stateful Polya trees correspond to mixtures of standard Polya trees, and the use of marginal likelihoods to determine posterior probabilities of states is equivalent to a Bayesian model averaging problem over the mixture components. Here we consider the case of a simple optional Polya tree (Wong and Ma, 2010), which corresponds to a mixture of finite Polya trees with different depths and partition structures. In the mixture representation we can identify two parts of the posterior computations: calculating the posterior for each finite Polya tree model, and calculating the posterior mixture probabilities over the models. In the stateful representation these correspond to the computation of the posteriors for the mass assignment parameters and the computation of the posterior probabilities of states, respectively. It is in the latter part that the trouble with marginal likelihoods arises, causing an excessive preference for the stopped state and resulting in blocky density estimates. The optional Polya tree as described by Wong and Ma (2010) uses Beta$(\frac{1}{2}, \frac{1}{2})$ priors for the mass assignment parameters, which strongly favors the stopped state when the observed data distribution is near 0.5.

We can apply the fractional Bayes factor methodology to this problem by using a fractional power of the likelihood to update the prior distributions for the mass assignment parameters without touching the probabilities of the states. In the mixture representation this corresponds to updating the prior of each finite Polya tree model without changing the prior distribution over models. The remainder of the data is then used to update both parts of the posterior in the standard way. The posteriors of the mass assignment parameters—or the posterior of each individual finite Polya tree model—is identical to the standard update under this scheme; only the posterior state probabilities, or the posterior distribution over finite Polya trees, differs.

Figure 3.5 shows a typical example of how the fractional Bayes factor approach affects the optional Polya tree. One thousand data points are simulated from a Beta$(10, 20)$ distribution and a optional Polya tree is fit with the fractional Bayes factor modification. Each panel corresponds to a different fraction of the likelihood used in the initial training set. A fraction of zero is equivalent to the standard optional Polya tree, and the blockiness of the posterior is evident. As the fraction of the likelihood used for training increases, the preference for the stopped state initially decreases. It eventually (when the training fraction reaches 1) reaches the prior probability of 0.5, following a curve like that shown in Figure 3.2. Each pane shows the L1 distance between the true density and the posterior mean density, defined as

$$d(f, g) = \frac{1}{2} \int |f(x) - g(x)| \, dx.$$

As can be seen, splitting the likelihood results in more accurate inference to a point, as the bias towards blockiness caused by Lindley's paradox is decreased. Eventually, however, the increased noise in the estimate becomes more important than the decreased bias, and larger training fractions become deleterious. Bias returns in a different form when the training fraction approaches one, as the posterior probabilities of states are then determined entirely by the prior. As in the simple case of testing a proportion, discussed above, maximal benefit is obtained with a fairly small fraction used for training.

A simulation study helps illustrate the benefit of the fractional training. Once again samples are drawn from a Beta(10,20) distribution, though the sample size is increased to 10,000. An optional Polya tree is fit to each sample with training fractions running from 0 to 0.2, and the L1 distance between the true density and the posterior mean is calculated as before. Figure 3.6 shows the mean L1 error as the training fraction varies, as well as error curves for 100 individual samples to

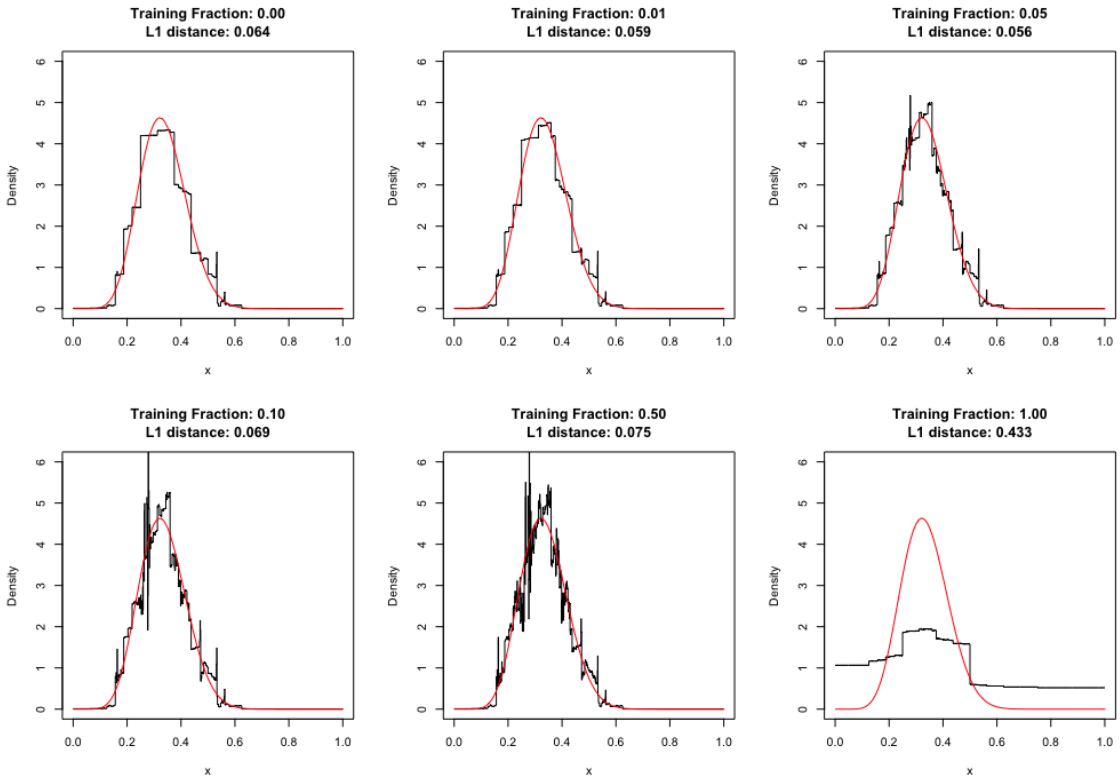

FIGURE 3.5: A typical example of the effect of the fractional Bayes factor approach on an optional Polya tree. Six training fractions are illustrated, with the true density in red and the posterior mean of the OPT in black.

illustrate variability of the error rate. For this simulation setting the average L1 error is optimized at a training fraction of about 0.03, with a decrease of 11% relative to the standard analysis. However, both the optimal training fraction and the benefit vary substantially between samples. Optimal training fractions vary from 0 to 0.2 (the largest value considered in this study) with a mean of 0.05, while the improvement in L1 error is as large as 33%, with a mean of 11%. A simulation study with the sample size set to 50,000 showed largely similar results, while a third simulation study with a sample size of 1,000 shows smaller benefits (mean improvement in L1 error of just 4.6%).

While substantial questions remain about the use of partial or fractional training sets to improve the fit of the prior distribution before carrying out Bayesian model evaluation procedures, these simulations demonstrate substantial gains in inferential accuracy when the method is applied to the optional Polya tree model. Similar results may be expected in other stateful Polya tree models where states correspond to parameter spaces with substantially different prior diffusion, such as the complete stoppage state in the adaptive (and hierarchical adaptive) Polya tree or the coupling states in multi-sample models designed for hypothesis testing (Ma and Wong, 2011; Soriano and Ma, 2017; Ma and Soriano, 2016).

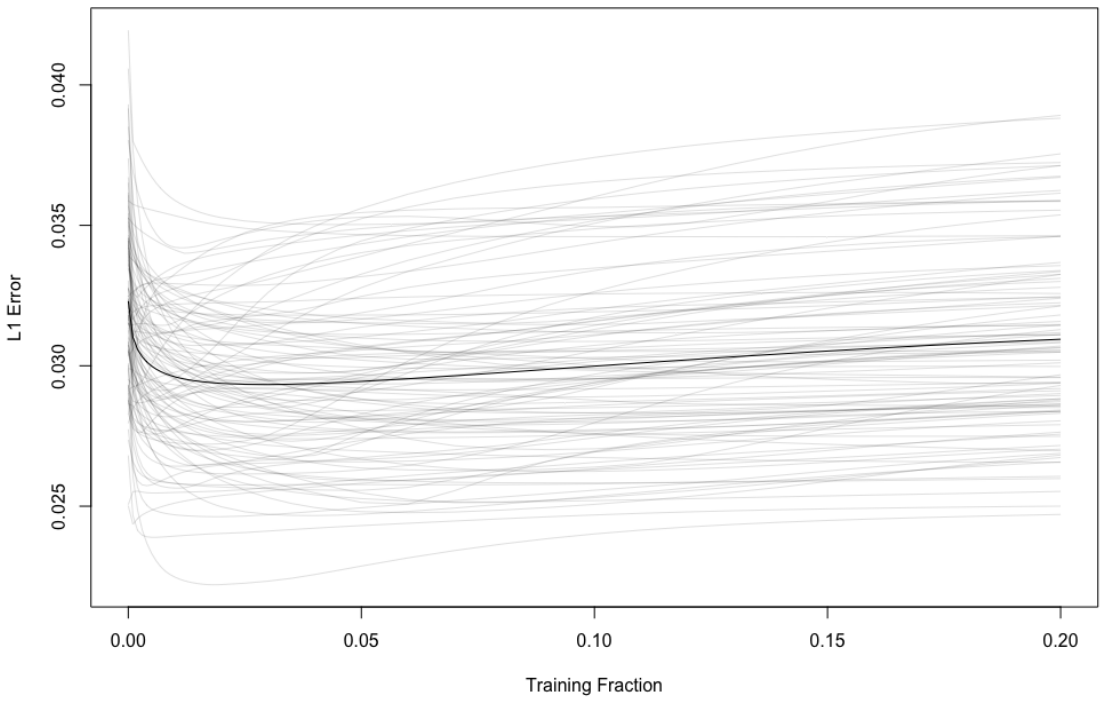

FIGURE 3.6: A simulation study shows how the L1 error between the true density and the posterior mean varies as the training fraction changes.

3.5 Discussion

Model evaluation is a central problem in Bayesian statistics, whether conducting formal hypothesis tests, averaging the results of several models to obtain better overall estimates, or comparing possible models during the model building process. While marginal likelihoods, posterior probabilities, and Bayes factors are often used in this context, problems with these approaches have long been recognized, spurring the development of alternative metrics by which to evaluate models; some examples include posterior predictive checks (Gelman et al., 1996), the Bayesian reference criterion of Bernardo et al. (1999), and the mixture representation of Kamary et al. (2014). These methods side step the difficulties associated with marginal-likelihood based methods but are conceptually unsatisfying: they lack the probabilistic interpretation that marginal likelihoods supply, and they fail to explain *why* marginal likelihoods fail.

This chapter answers the latter question and suggests an approach to resolve the problem. Partial and fractional likelihoods are rooted in a long statistical tradition of training/test data splits and retain the probabilistic interpretation of methods based on the marginal likelihood, while improving inferential accuracy in both simple and complex models. The partial likelihood approach has a natural interpretation in terms of sequential analysis; the fractional approach lacks this interpretability, but offers a computationally inexpensive approximation to averaging over a large number of training/test splits with the partial likelihood. By avoiding excessive penalization for priors not fitting the data, these methods promise more accurate statistical inference with all the advantages of marginal likelihood-based methods.

4

Conclusions

The stateful Polya tree paradigm offers a powerful approach to nonparametric modeling beyond simple density estimates. The first application of stateful Polya trees in Wong and Ma (2010) introduced a powerful new approach to analysis of multivariate data using Polya trees, offering a more parsimonious model than the traditional multivariate Polya tree. As demonstrated by Ma and Wong (2011); Ma and Soriano (2016); Soriano and Ma (2017), stateful Polya trees are also capable of handling two or more samples in a hypothesis testing framework, identifying local differences between the samples. Ma (2017) introduced a stateful Polya tree that allows the scale and smoothness of distribution features to be learned from the data differentially across the sample space by placing a fully nonparametric prior on the Polya tree's concentration parameter, rather than fixing it in the prior.

Chapter 2 showed how the stateful Polya tree framework can be used to model multiple samples with an emphasis on estimation and information sharing rather than hypothesis testing, identifying common structure and idiosyncratic features of each sample. We show how the states capture information about the dispersion between sample densities at different locations in the sample space, and how this

information can be summarized in a dispersion function.

Chapter 3 considers issues with the traditional use of marginal likelihoods in Bayesian hypothesis testing. In the context of stateful Polya trees this problem arises, for example, in deciding whether to stop cutting in the optional Polya tree (Wong and Ma, 2010) or the multi-resolution scanning model (Ma and Soriano, 2016), in the multi-sample hypothesis tests of Ma and Wong (2011) and Soriano and Ma (2017), and in the computation of shrinkage state probabilities in (Ma, 2017) and the HAPT model described in Chapter 2. However, the question is a foundational one with implications far beyond stateful Polya tree models.

Stateful Polya trees show potential to tackle a number of other problems which have not yet been fully explored. For example, Paddock (2002) describes a Gibbs sampler to impute missing data in a Polya tree model. The parameter space for this model is extremely high-dimensional, and there may be considerable concern about the mixing and convergence of a Markov chain. A simple analysis shows that given a truncated Polya tree prior and data observed with missing values, the posterior is a mixture of truncated Polya trees with a large but finite number of mixture components. Neath (2003) considered the case of censored data and demonstrated computation of the posterior mixture distribution with a very small dataset and a shallow tree, but the number of components becomes intractable as the size of the dataset and the depth of the tree grow. The Gibbs sampler described in Paddock (2002) avoids enumerating the mixture by sampling random components, subject to concerns over accuracy of the MCMC sampler.

Given the representations of stateful Polya trees as mixtures of Polya trees discussed in the introduction, it is perhaps not surprising that we can describe a stateful Polya tree which results in the same posterior distribution, with states representing the locations of the unobserved values at each node of the tree. Using a stateful Polya tree allows posterior computation using a recursive algorithm, avoiding the

intractably large number of components in the mixture representation. This model is the subject of current work.

Bibliography

Anderson, T. W. (1962), "On the Distribution of the Two-Sample Cramer-von Mises Criterion," *The Annals of Mathematical Statistics*, 33, 1148–1159.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013), "Modeling with Normalized Random Measure Mixture Models," *Statistical Science*, 28, 313–334.

Bayarri, M. J. and García-Donato, G. (2008), "Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70, 981–1003.

Berger, J. and Guglielmi, A. (2001), "Bayesian and Conditional Frequentist Testing of a Parametric Model versus Nonparametric Alternatives," *Journal of the American Statistical Association*, 96, 174–184.

Berger, J. O. and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bernardo, J. M. and Rueda, R. (2002), "Bayesian Hypothesis Testing: A Reference Approach," *International Statistical Review*, 70, 351–372.

Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith (eds, A. F. M., and Bernardo, J. t. M. (1999), *Nested Hypothesis Testing: The Bayesian Reference Criterion, Bayesian Statistics 6*.

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017), "Bayesian Prediction with Multiple-Samples Information," *Journal of Multivariate Analysis*, 156, 18–28.

CDC (2017), "CDC WONDER," `https://wonder.cdc.gov/`, Accessed: 2017-04-14.

Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., Altshuller, Y. M., Frohman, M. A., Kraner, S. D., and Mandel, G. (1995), "REST: A Mammalian Silencer Protein That Restricts Sodium Channel Gene Expression to Neurons," *Cell*, 80, 949–957.

Christensen, J. and Ma, L. (2017), "A Bayesian Hierarchical Model for Related Densities Using Polya Trees," *arXiv:1710.01702 [stat]*.

Conigliani, C. and O'Hagan, A. (2000), "Sensitivity of the Fractional Bayes Factor to Prior Distributions," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28, 343–352.

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998), "Wavelet-Based Statistical Signal Processing Using Hidden Markov Models," *IEEE Transactions on Signal Processing*, 46, 886–902.

De Santis, F. and Spezzaferri, F. (1997), "Alternative Bayes Factors for Model Selection," *Canadian Journal of Statistics*, 25, 503–515.

de Santis, F. and Spezzaferri, F. (1999), "Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach," *International Statistical Review / Revue Internationale de Statistique*, 67, 267–286.

ENCODE Project Consortium (2012), "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature*, 489, 57–74.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Ferguson, T. S. (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Freedman, D. A. (1963), "On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case," *The Annals of Mathematical Statistics*, 34, 1386–1403.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "POSTERIOR PREDICTIVE ASSESSMENT OF MODEL FITNESS VIA REALIZED DISCREPANCIES," *Statistica Sinica*, 6, 733–760.

Hanson, T. and Johnson, W. O. (2002), "Modeling Regression Error with a Mixture of Polya Trees," *Journal of the American Statistical Association*, 97, 1020–1033.

Hanson, T. E. (2006), "Inference for Mixtures of Finite Polya Tree Models," *Journal of the American Statistical Association*, 101, 1548–1565.

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015), "Two-sample Bayesian Nonparametric Hypothesis Testing," *Bayesian Anal.*, 10, 297–320.

Jefferys, W. H. (1990), "Bayesian Analysis of Random Event Generator Data," *Journal of Scientific Exploration*, 4, 153–169.

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014), "Testing Hypotheses via a Mixture Estimation Model," *arXiv:1412.2044 [stat]*.

Kolmogorov, A. (1933), "Sulla Determinazione Empirica Di Una Legge Di Distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.

Kraft, C. H. (1964), "A Class of Distribution Function Processes Which Have Derivatives," *Journal of Applied Probability*, 1, 385–388.

Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modelling," *Ann. Statist.*, 20, 1222–1235.

Lavine, M. (1994), "More Aspects of Polya Tree Distributions for Statistical Modelling," *Ann. Statist.*, 22, 1161–1176.

Lindley, D. V. (1957), "A STATISTICAL PARADOX," *Biometrika*, 44, 187–192.

Ma, L. (2017), "Adaptive Shrinkage in Pólya Tree Type Models," *Bayesian Analysis*, 12, 779–805.

Ma, L. and Soriano, J. (2016), "Analysis of Distributional Variation through Multi-Scale Beta-Binomial Modeling," *arXiv:1604.01443 [stat]*.

Ma, L. and Wong, W. H. (2011), "Coupling Optional Pólya Trees and the Two Sample Problem," *Journal of the American Statistical Association*, 106, 1553–1565.

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55, Alexandria, Virginia. Virginia: American Statistical Association; 1999.

MacEachern, S. N. (2008), "Discussion of "The nested Dirichlet process" by A.E. Gelfand, D.B. Dunson and A. Rodriguez," *Journal of the American Statistical Association*, 103, 1149–1151.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992), "Polya Trees and Random Distributions," *Ann. Statist.*, 20, 1203–1221.

Muliere, P. and Walker, S. (1997), "A Bayesian Non-Parametric Approach to Survival Analysis Using Polya Trees," *Scandinavian Journal of Statistics*, 24, 331–340.

Müller, P., Quintana, F., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 735–749.

Neath, A. A. (2003), "Polya Tree Distributions for Statistical Modeling of Censored Data," https://www.hindawi.com/journals/ads/2003/745230/abs/.

Nieto-Barajas, L. E. and Müller, P. (2012), "Rubbery Polya Tree," *Scandinavian Journal of Statistics*, 39, 166–184.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 99–138.

O'Hagan, A. (1997), "Properties of Intrinsic and Fractional Bayes Factors," *Test*, 6, 101–118.

Paddock, S. M. (2002), "Bayesian Nonparametric Multiple Imputation of Partially Observed Data with Ignorable Nonresponse," *Biometrika*, 89, 529–538.

Paddock, S. M., Ruggeri, F., Lavine, M., and West, M. (2003), "RANDOMIZED POLYA TREE MODELS FOR NONPARAMETRIC BAYESIAN INFERENCE," *Statistica Sinica*, 13, 443–460.

Pericchi, L. R. (2005), "Model Selection and Hypothesis Testing Based on Objective Probabilities and Bayes Factors," in *Handbook of Statistics*, eds. D. K. Dey and C. R. Rao, vol. 25 of *Bayesian Thinking*, pp. 115–149, Elsevier.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008), "The Nested Dirichlet Process," *Journal of the American Statistical Association*, 103, 1131–1154.

Song, L. and Crawford, G. E. (2010), "DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells," *Cold Spring Harbor protocols*, 2010, pdb.prot5384.

Soriano, J. and Ma, L. (2017), "Probabilistic Multi-Resolution Scanning for Two-Sample Differences," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 547–572.

Teh, Y. W. (2006), "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 985–992, Stroudsburg, PA, USA, Association for Computational Linguistics.

Teh, Y. W. and Jordan, M. I. (2010), *Hierarchical Bayesian nonparametric models with applications*, pp. 158–207, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Walker, S. G. and Mallick, B. K. (1997), "Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 845–860.

Westenberg, J. (1948), "Significance Test for Median and Interquartile Range in Samples from Continuous Populations of Any Form," *Proceedings Koninklijke Nederlandse Akademie van Wetenshappen*, 51, 252–261.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1, 80–83.

Wong, W. H. and Ma, L. (2010), "Optional Pólya tree and Bayesian inference," *Ann. Statist.*, 38, 1433–1459.

Zhao, L. and Hanson, T. E. (2011), "Spatially Dependent Polya Tree Modeling for Survival Data," *Biometrics*, 67, 391–403.

Biography

Jonathan Casey Christensen was born in Fairfax County, Virginia, October 11 1987. He earned a BS Mathematics and MS Statistics from Brigham Young University (2012), an MS Statistics from Duke University (2015), and a PhD Statistics from Duke University (2017).