Coupling optional Polya trees and the two sample problem

Li Ma* Wing H. Wong†

November 5, 2010

Abstract

Testing and characterizing the difference between two data samples is of fundamental interest in statistics. Existing methods such as Kolmogorov-Smirnov and Cramer-von-Mises tests do not scale well as the dimensionality increases and provides no easy way to characterize the difference should it exist. In this work, we propose a theoretical framework for inference that addresses these challenges in the form of a prior for Bayesian nonparametric analysis. The new prior is constructed based on a random-partition-and-assignment procedure similar to the one that defines the standard optional Polya tree distribution, but has the ability to generate multiple random distributions jointly. These random probability distributions are allowed to "couple", that is to have the same conditional distribution, on subsets of the state space. We show that this "coupling optional Polya tree" prior provides a convenient and effective way for both the testing of two sample difference and the learning of the underlying structure of the difference. In addition, we discuss some practical issues in the computational implementation of this prior and provide several numerical examples to demonstrate its work.

1 Introduction

Two sample comparison is a fundamental problem in statistics. With two samples of data at hand, one often wants to answer the question—"Are these two samples different?" In slightly more statistical language, one is interseted in testing the null hypothesis that the two data samples were generated from the same distribution. In addition, in the presence of evidence for deviation between the two samples, one often hopes to learn the structure of such difference in order to understand, for example, what factors could have played a role in causing the difference. Hence two sample comparison is interesting both as a hypothesis testing problem and as a data mining problem. In this work, we consider the problem from both aspects, and develop a Bayesian nonparametric approach that can serve both the testing and the learning purposes.

*Department of Statistics, 390 Serra Mall, Stanford, CA 94305 / EM: ma2@stanford.edu
†Department of Statistics, 390 Serra Mall, Stanford, CA 94305 / EM: whwong@stanford.edu

Nonparametric hypothesis testing for two sample difference has a long history and rich literature, and many methods have been proposed. Some well-known examples include Wilcoxon test [12, p.243], Kolmogorov-Smirnov test [6, pp. 392–394] and Cramer-von-Mises test [1]. Recently, this problem has also been investigated from a Bayesian nonparametric perspective using the Polya tree prior [10].

Despite the success of these existing testing methods for one-dimensional problems, two sample comparison in multi-dimensional spaces remains a challenging task. A basic idea for many existing methods is to estimate the two underlying distributions, and then use a distance metric to measure the dissimilarity between the two estimates. Tests such as Kolmogorov-Smirnov (K-S) and Cramer-von-Mises (CvM) fall into this category. However, reliably characterizing distributions in multi-dimensional problems, if computationally feasible at all, often requires a prohibitively large number of data points. With even just a moderate number of dimensions, the estimated distributional distance is often highly variable or biased. This is true even when the underlying difference is structurally very simple and can be accounted for by a relatively small number of dimensions in the space. This so-called "curse of dimensionality" demonstrates itself in the Bayesian context as well.

One general approach to dealing with the curse of dimensionality when characterizing distributions in a multi-dimensional space is to learn from the data a partition of the space that best reflects the underlying structure of the distribution(s). A good partition of the space overcomes the sparsity of the data by placing true neighbors together, and it reduces computational burden by allowing one to focus on the relevant blocks in the space. Hence it can be very helpful in multi-dimensional, and especially high-dimensional, settings to incorporate the learning of a representative partition of the space into the inference procedure. Wong and Ma [18] adopted this idea and introduced the optional Polya tree (OPT) prior as such a method under the Bayesian nonparametric framework. Through optional stopping and randomized splitting of the state space, a recursive partitioning procedure is incorporated into the parametrization of this prior, thereby allowing the data to suggest parsimonious divisions of the space. Althought the OPT prior deals with only one data sample, similar ideas can also be utilized for problems involving more than one sample as will be demonstrated in this work.

Besides the difficulty in handling multidimensional problems, existing nonparametric methods for two sample comparison are also unsatisfactory in that they provide no easy way to learn the underlying structure of the difference should it exist. Tests such as K-S and CvM provide statistics with which to test for the existence of a difference, but does not allow one to characterize the difference—for example what variables are involved in the difference and how. One has to resort to methods such as logistic regression that rely on strong modelling assumptions to investigate such structure. Similarly, a Bayes factor computed using nonparametric priors such as Dirichlet process mixture and the Polya tree prior also sheds no light on where the evidence for difference has arisen.

In this work, we introduce a new prior called "coupling optional Polya tree" (co-OPT) designed for Bayesian nonparametric inference on the two sample problem. This new prior *jointly* generates two random distributions through a random-partitioning-and-assignment

procedure similar to the one that gives rise to the OPT prior [18]. The co-OPT framework allows both hypothesis testing on the null hypothesis and posterior learning of the distributional difference in terms of a partition of the space that "best" reflects the difference structure. The ability to make posterior inference on a partition of the space also enhances the testing power for multi-dimensional problems.

This paper is organized as follows. In Section 2 we review the construction of the OPT distribution. In Section 3 we generalize the definition of the OPT distribution by replacing the "uniform base measure" (defined later) with a general absolutely continuous distribution, and show that this generalized prior can be used for investigating the goodness-of-fit of the data to the base distribution. In Section 4 we introduce the co-OPT prior and show how Bayesian inference can be carried out using this prior. In addition, we discuss the practical issues in implementing inference using this prior. In Section 5 we provide several numerical examples to illustrate inference on the two sample comparison problem using this prior. Then in Section 6 we present a method for inferring two common distributional distances, L_1 and Hellinger, between the two sample distributions using a co-OPT prior and provide two more numerical examples. Section 7 concludes with a few remarks.

We close this introduction with a few words on the recent development in the Bayesian nonparametric literature on related topics. In the past 15 years, several methods have been proposed for testing the one sample goodness-of-fit, in particular, for non-parametric alternatives against a parametric null. For some examples see [7, 5, 4, 3, 9, 14, 17]. As for two sample comparison, Holmes *et. al.* [10] introduced a way to compute the Bayes factor for testing the null through the marginal likelihood of the data with Polya tree priors. Under the null, they model the two samples to have come from a single random measure distributed as a Polya tree, while under the alternative from two separate Polya tree distributions. In contrast, our new prior allows the two distributions to be generated *jointly* through one prior even when they are different. It is this joint generation that allows both the testing of the difference and the learning of the structure simultaneously. There are other approaches to joint modeling of multiple distributions in the Bayesian nonparametric literature. For example, one idea is to use "dependent Dirichlet processes" [13]. For some notable examples developed in this framework see [15, 16, 8], among many others. Compared to these methods based on Dirichlet processes, our method, based on the optional Polya tree, allows the resolution of the inference to be adaptive to the data structure and handles the sparsity in multidimensional settings using random partitioning [18]. Moreover, our method allows direct inference on the distributional difference without relying on inferring the two distributions *per se*, making it particularly suited for comparison across multiple samples. This point will be further discussed in Section 4.2 and illustrated in the examples given in Sections 5 and 6.

2 Optional Polya trees and Bayesian inference

Wong and Ma [18] introduced the optional Polya tree (OPT) distribution as an extension to the Polya tree prior that allows optional stopping and randomized partitioning of the state

space Ω, where Ω is either finite or a rectangle in an Euclidean space. One can think of this prior as a procedure for generating random probability measures on Ω that consists of two components—(1) random partitioning of the space and (2) random probability assignment into the parts of the space produced by the partitioning.

We first review how the OPT prior randomly partitions the space. Let $\mathcal{R}$ denote a partition rule function which, for any subset A of Ω, defines a number of ways to partition A into a finite number of smaller sets. For example, for $\Omega = [0, 1]$, the diadic split rule $\mathcal{R}$ is that $\mathcal{R}(A)$=$\{$ cutting A in the middle of its range $\}$ if A is an non-empty interval and $= \emptyset$ otherwise. We call a rule function $\mathcal{R}$ finite if $\forall A \subset \Omega$, we have $M(A) := |\mathcal{R}(A)| < \infty$, in which case we can say A can be divided in $M(A)$ ways in the usual sense. In the rest of the paper, we will only consider finite partition rules. Let $K^j(A)$ be the number of children specified by the jth way to partition A under $\mathcal{R}(A)$, and let A_i^j denote the ith child set of A in that way of partitioning. We can write $\mathcal{R}(A)$ as

$$\mathcal{R}(A) = \{\{A_1^1, A_2^1, \dots A_{K^1}^1\}, \{A_1^2, A_2^2, \dots, A_{K^2}^2\}, \dots \{A_1^M, A_2^M, \dots A_{K^M}^M\}\} = \left\{\{A_i^j\}_{i=1}^{K^j}\right\}_{j=1}^{M},$$

where for simplicity we suppressed notation by writing M for $M(A)$ and K for $K(A)$.

A partition rule function $\mathcal{R}$ does not specify any particular partition on Ω but rather a collection of possible partitions over which one can draw random samples. The OPT prior samples from this collection of partitions in the following sequential way. Starting from the whole space $A = \Omega$. If $M(A) = 0$, then A is not divisible under $\mathcal{R}$ and we call A an atom (set). In this case the partitioning of A is completed. If $M(A) > 0$, that is, A is divisible, then a Bernoulli($\rho(A)$) random variable $S(A)$ is drawn. If $S(A) = 1$, we stop partitioning A. Hence $S(A)$ is called the stopping variable for A, and $\rho(A)$ the stopping probability. If $S(A) = 0$, A is divided in the $J(A)$th of the $M(A)$ available ways for partitioning A under $\mathcal{R}(A)$, where $J(A)$ is a random variable taking values $1, 2, \dots, M(A)$ with probabilities $\lambda_1(A), \lambda_2(A), \dots, \lambda_{M(A)}(A)$ respectively, and $\sum_{j=1}^{M(A)} \lambda_j(A) = 1$. $J(A)$ is hence called the (partition) selector variable, and $\boldsymbol{\lambda}(A) = (\lambda_1(A), \lambda_2(A), \dots, \lambda_{M(A)}(A))$ the (partition) selector probabilities. If $J(A) = j$, we partition A into $\{A_1^j, A_2^j, \dots A_{K^j(A)}^j\}$, and then apply the same procedure to each of the children. In addition, if A is reached from Ω after k steps (or levels) of such nested recursive partitioning (NRP), then we say that A_i^j is reached after $k + 1$ steps (or levels) of NRP. (To complete this inductive definition, we say that the space Ω is reached after 0 steps of NRP.)

The first question that naturally arises is whether this sequential procedure will eventually "stop" and produce a well defined partition on Ω. Given that the stopping probability $\rho(A) > \delta$ for some δ and all A, this is indeed true in the following sense. If we let μ be the natural measure on Ω—the Lebesgue measure if Ω is a rectangle in an Euclidean space or the counting measure if Ω is finite, then $\mu(T_1^k) \to 0$ with probability 1, where T_1^k is the part of Ω that is still not stopped after k steps of NRP. In other words, the partitioning procedure will stop almost everywhere on Ω.

The second component of the OPT prior is random probability assignment. The prior assigns probability mass into the randomly generated parts of the space in the following

manner. Starting from $A = \Omega$, assign $Q(A) = 1$ total probability to A. If A is stopped or is an atom, then let the conditional distribution within A be uniform. That is, $Q(\cdot|A) = u(\cdot|A)$, where u denotes the uniform density (w.r.t. μ) and this completes the probability assignment on A. If instead A has children $\{A_1^j, A_2^j, \ldots A_{K^j(A)}^j\}$, (this occurs when $S(A) = 0$ and $J(A) = j$,) a random vector $(\theta_1^j(A), \theta_2^j(A), \ldots, \theta_{K^j(A)}^j(A))$ on the $K^j(A) - 1$ dimensional simplex is drawn from a Dirichlet$(\alpha_1^j(A), \alpha_2^j(A), \ldots, \alpha_{K^j(A)}^j(A))$ distribution, and we assign to each child A_i^j probability mass $Q(A_i^j) = Q(A)\theta_i^j(A)$. We call $\boldsymbol{\theta}^j(A) = (\theta_1^j(A), \theta_2^j(A), \ldots, \theta_{K^j(A)}^j(A))$ the (probability) assignment vector, and $\boldsymbol{\alpha}^j(A) = (\alpha_1^j(A), \alpha_2^j(A), \ldots, \alpha_{K^j(A)}^j(A))$ the pseudo-count parameters. Then we go to the next level and assign probability mass within each of the children in the same way.

Theorem 1 in [18] shows that if $\rho(A) > \delta$ for some $\delta > 0$ and all A, then with probability 1 this random partitioning and assignment procedure will give rise to a probability measure Q on Ω that is absolutely continuous with respect to μ. This random measure Q is said to have an OPT distribution with (partition rule $\mathcal{R}$ and) parameters ρ, $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, which can be written as $OPT(\mathcal{R}; \rho, \boldsymbol{\lambda}, \boldsymbol{\alpha})$. In addition, Wong and Ma [18] also show that under mild conditions, this prior has large support—any L_1 neighborhood of an absolutely continuous distribution (w.r.t. μ) on Ω has positive prior probability.

Two key features of the prior are demonstrated from the above constructive description. The first is self-similarity. If a set A is reached as a node during the procedure, then the continuing partitioning and assignment *within* A, which specifies the conditional distribution on A, is just an OPT procedure with $\Omega = A$. The second feature is the prior's implicit hierarchical structure. To see this, we note that the random distribution that arises from such a prior is completely determined by the partition and assignment variables S, J, and $\boldsymbol{\theta}$, while the prior parameters ρ, $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ specify the distributions of these "middle" variables.

These two features allow one to write down a recursive formula for the likelihood under a random distribution arising from such a prior. To see this, first let Q (with density q) be a distribution arising from an $OPT(\mathcal{R}; \rho, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ distribution, and for $A \subset \Omega$, let $q(\cdot|A)$ be the conditional density on A. Let S, J, and $\boldsymbol{\theta}$ be the corresponding partition and assignment variables for Q (or q). Suppose one has n i.i.d. observations, x_1, x_2, …, x_n, on Ω from $q(\cdot|\Omega)$. Define

$$\boldsymbol{x}(A) = \{x_1, x_2, \ldots, x_n\} \cap A,$$

the observations falling in A, and $n(A) = |\boldsymbol{x}(A)|$, the number of observations in A. Then for any A that is a node reached in the recursive partitioning process determined by the S and J variables, the likelihood of observing $\boldsymbol{x}(A)$ conditional on A is

$$q\left(\boldsymbol{x}(A)|A\right) = Su\left(\boldsymbol{x}(A)|A\right) + (1 - S)\left(\prod_{i=1}^{K^J}\left(\theta_i^J\right)^{n(A_i^J)}\right)\left(\prod_{i=1}^{K^J} q\left(\boldsymbol{x}\left(A_i^J\right)\left|A_i^J\right.\right)\right), \tag{2.1}$$

where $u(\boldsymbol{x}(A)|A) = \frac{1}{\mu(A)^{n(A)}}$ is the likelihood under the uniform distribution on A, $S = S(A)$, $J = J(A)$, $K^J = K^{J(A)}(A)$, and $\theta_i^J = \theta_i^{J(A)}(A)$. (Note that for this formula to hold we need to define $q\left(\emptyset|A\right) := 1$.) From now on we will always suppress the "(A)" notation for the

random variables and the parameters where this adds no confusion. Similarly, we will use $q(\boldsymbol{x}|A)$ and $u(\boldsymbol{x}|A)$ to mean $q(\boldsymbol{x}(A)|A)$ and $u(\boldsymbol{x}(A)|A)$, respectively.

Integrating out S, J, and $\boldsymbol{\theta}$ in (2.1), we get the corresponding recursive representation of the marginal likelihood

$$P(\boldsymbol{x}|A) = \rho u(\boldsymbol{x}|A) + (1-\rho)\sum_{j=1}^{M} \lambda_j \frac{D(\boldsymbol{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} P\left(\boldsymbol{x}|A_i^j\right), \tag{2.2}$$

where $P(\boldsymbol{x}|A) = P(\boldsymbol{x}(A)|A)$, $\boldsymbol{n}^j = \boldsymbol{n}^j(A) = (n(A_1^j), n(A_2^j), \ldots, n(A_{K^j(A)}^j))$, and $D(\boldsymbol{t}) = \Gamma(t_1)\ldots\Gamma(t_k)/\Gamma(t_1+\cdots+t_k)$. This provides a recipe for computing the marginal likelihood conditional on A, $P(\boldsymbol{x}|A)$, for all potential tree nodes A determined by $\mathcal{R}$.

The final result we review in the section is the conjugacy of the OPT prior. More specifically, given the i.i.d. observations $\boldsymbol{x}$, the posterior distribution of Q is again an OPT distribution with

1. Stopping probability:
$$\rho(A|\boldsymbol{x}) = \rho(A)u(\boldsymbol{x}|A)/P(\boldsymbol{x}|A)$$

2. Selection probabilities:
$$\lambda_j(A|\boldsymbol{x}) \propto \lambda_j(A)\frac{D(\boldsymbol{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} P\left(\boldsymbol{x}|A_i^j\right) \qquad \text{for } j = 1, \ldots, M(A)$$

3. Probability assignment pseudo-counts:
$$\alpha_i^j(A|\boldsymbol{x}) = \alpha_i^j(A) + n(A_i^j)$$
for $j = 1, \ldots, M(A)$ and $i = 1, 2, \ldots, K^j(A)$

where again A is any potential node determined by the partition rule function $\mathcal{R}$ on Ω.

3 Optional Polya trees with general base measures

In the constructive procedure for an OPT distribution described above, whenever a node A is stopped, the conditional distribution within it is generated from that of a baseline distribution, namely the uniform $u(\cdot|A)$. For this reason, we say that the collection of conditional uniform distributions, $\{u(\cdot|A) : A \text{ is a potential node under } \mathcal{R}\}$, are the *local* base measures. With uniform local base measures, the stopping probability ρ for a region A represents the probability that the distribution is "flat" within A. Accordingly, the posterior OPT concentrates probability mass around partitions that best captures the "non-flatness" in the density of the data distribution. Such a partitioning criterion is most natural in the context of density estimation.

One can extend the original OPT construction by adopting different local base measures or stopping criteria for the nodes. More specifically, we can replace $u(\cdot|A)$ with any absolutely continuous measure $m^A(\cdot)$ on node A in the probability assignment step. That is, when a tree node A is stopped, we let the conditional distribution in A be $m^A(\cdot)$. With this generalization, the recursive constructive procedure for the OPT distribution and the recipe for Bayesian inference described in the previous section still follow through.

One choice of the m^A measures is of particular interest. Specifically, we can let $m^A(\cdot) = q_0(\cdot|A)$ for some absolutely continuous distribution Q_0 with density q_0 on Ω. For this special case, we have the following definition.

Definition 1. The random probability measure Q that arises from the random-partitioning-and-assignment (RPAA) procedure described in the previous section, with u replaced by q_0, the density (w.r.t. μ) of an absolutely continuous distribution Q_0, is said to have an optional Polya tree distribution on $\mathcal{R}$ with parameters $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, ρ, and (global) base measure (or distribution) Q_0. We denote this distribution by $OPT(\mathcal{R}; \boldsymbol{\lambda}, \boldsymbol{\alpha}, \rho; Q_0)$.

The next theorem shows that by choosing an appropriate partitioning rule $\mathcal{R}$ and/or suitable pseudocount parameters $\boldsymbol{\alpha}$, one can enforce the random distribution Q to "center around" the base measure Q_0.

Theorem 1. *If $Q \sim OPT(\mathcal{R}; \rho, \boldsymbol{\lambda}, \boldsymbol{\alpha}; Q_0)$, where $\delta < \rho(A)$ for some δ and all potential tree nodes A, then $\forall$ Borel set B,*

$$EQ(B) = Q_0(B),$$

provided that for all A, $j = 1, 2, \ldots, M(A)$ and $i = 1, 2, \ldots, K^j(A)$, we have

$$\alpha_i^j(A) / \sum_{h=1}^{K^j(A)} \alpha_h^j(A) = Q_0(A_i^j)/Q_0(A).$$

Proof. Consider the RPAA procedure described in the previous section with the uniform base distribution u replaced by q_0. So under this new procedure of generating a random measure Q, whenever a region A gets stopped, the conditional distribution of Q within A is set to be $Q_0(\cdot|A)$. Let $Q^{(k)}$ be the corresponding random distribution that is forced to stop after k levels of nested partitioning. In other words, for all non-stopped nodes A reached after k levels of nested partitioning, we stop dividing A regardless of the stopping variable $S(A)$ and force a conditional distribution $Q_0(\cdot|A)$ on it to obtain $Q^{(k)}$. (For more detail see the proof of Theorem 1 in [18].)

We first show that if $\alpha_i^j(A)/\sum_{h=1}^{K^j(A)} \alpha_h^j(A) = Q_0(A_i^j)/Q_0(A)$, then $EQ^{(k)}(B) = Q_0(B)$ for all k. For $k \geq 0$, let $\mathcal{J}^{(k)}$ be the collection of all partition random variables S and J drawn in the first k levels of partitioning, and let $\mathcal{A}(\mathcal{J}^{(k)})$ be the collection of all leaf nodes after k levels of random partitioning—those are the nodes that are either just reached in the kth step or are reached earlier but stopped. We prove by induction that $E\left(Q^{(k)}(B)|\mathcal{J}^{(k)}\right) = Q_0(B)$.

For $k=0$, $\mathcal{J}^{(k)}=\emptyset$, $\mathcal{A}(\mathcal{J}^{(k)})=\{\Omega\}$, and $Q^{(0)}=Q_0$ and so $E\left(Q^{(k)}(B)|\mathcal{J}^{(k)}\right)=Q_0(B)$ holds trivially. Now for $k\geq 1$, suppose this holds true for $1,2,\ldots,k-1$. By construction,

$$Q^{(k)}(B)=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}Q^{(k)}(A)\frac{Q_0(B\cap A)}{Q_0(A)}.$$

Let $A^p\in\mathcal{A}(\mathcal{J}^{(k-1)})$ be the parent node of A, that is, the node whose division gives rise to A. Then by the condition that $\alpha_i^j(A)/\sum_{h=1}^{K^j(A)}\alpha_h^j(A)=Q_0(A_i^j)/Q_0(A)$, we have

$$E(Q^{(k)}(A)/Q^{(k)}(A^p)|\mathcal{J}^{(k)})=Q_0(A)/Q_0(A^p),$$

and so

$$\begin{aligned}
E\left(Q^{(k)}(B)|\mathcal{J}^{(k)}\right)&=E\left(\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}Q^{(k)}(A)\frac{Q_0(B\cap A)}{Q_0(A)}\Big|\mathcal{J}^{(k)}\right)\\
&=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}\frac{Q_0(B\cap A)}{Q_0(A)}E\left(Q^{(k)}(A)\Big|\mathcal{J}^{(k)}\right)\\
&=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}\frac{Q_0(B\cap A)}{Q_0(A)}E\left(\frac{Q^{(k)}(A)}{Q^{(k)}(A^p)}Q^{(k)}(A^p)\Big|\mathcal{J}^{(k)}\right)\\
&=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}\frac{Q_0(B\cap A)}{Q_0(A)}\frac{Q_0(A)}{Q_0(A^p)}E\left(Q^{(k)}(A^p)\Big|\mathcal{J}^{(k)}\right)\\
&=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}\frac{Q_0(B\cap A)}{Q_0(A^p)}E\left(Q^{(k-1)}(A^p)\Big|\mathcal{J}^{(k)}\right)\\
&=\sum_{A\in\mathcal{A}(\mathcal{J}^{(k)})}Q_0(B\cap A)=Q_0(B).
\end{aligned}$$

This shows that $E\left(Q^{(k)}(B)|\mathcal{J}^{(k)}\right)=Q_0(B)$ and thus $EQ^{(k)}(B)=Q_0(B)$ for all k. But since $|Q^{(k)}(B)-Q(B)|\to 0$ a.s. (see the proof of Theorem 1 in [18]), by bounded convergence theorem, we have $E|Q^{(k)}(B)-Q(B)|\to 0$, and so $EQ(B)=Q_0(B)$. □

Remark: If we have equal pseudocounts, that is, $\alpha_1^j(A)=\alpha_2^j(A)=\cdots=\alpha_{K^j(A)}^j(A)$ for all potential nodes A and all j, then the condition for the theorem becomes $Q_0(A_i^j)/Q_0(A)=1/K^j(A)$. Therefore one can choose a partition rule $\mathcal{R}$ on Ω based on the base measure to center the prior.

Bayesian inference using the OPT prior with general base measures can be carried out just as before. More specifically, the recursive likelihood equation (2.1) becomes

$$q\left(\boldsymbol{x}|A\right)=Sq_0\left(\boldsymbol{x}|A\right)+(1-S)\left(\prod_{i=1}^{K^J}\left(\theta_i^J\right)^{n(A_i^J)}\right)\left(\prod_{i=1}^{K^J}q\left(\boldsymbol{x}\Big|A_i^J\right)\right),\tag{3.1}$$

and (2.2) becomes

$$P(\boldsymbol{x}|A) = \rho\, q_0(\boldsymbol{x}|A) + (1-\rho) \sum_{j=1}^{M} \lambda_j \frac{D(\boldsymbol{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} P\left(\boldsymbol{x}|A_i^j\right) \tag{3.2}$$

where $q_0(\boldsymbol{x}|A) = \prod_{x \in \boldsymbol{x}(A)} q_0(x|A)$, the likelihood of observing $\boldsymbol{x}(A)$ on A under $q_0(\cdot|A)$. The posterior parameter values are also same as before with u replaced by q_0.

An important fact is that a random distribution with this prior has positive probability to be *exactly* the same as the base distribution, which is also the mean of the OPT under the condition given in Theorem 1. Therefore, one can think of the inferential procedure for the OPT prior as a sequence of recursive comparison steps to the base measure. More specifically, the partitioning decision on each node A is determined by comparing the conditional likelihood of the data within A under Q_0 to the composite of $M(A)$ composite alternatives. The partition of each node A stops when the observations in A "fits" the structure of the base measure, and the posterior values of the partitioning variables capture the discrepancy, if any, between the data and the base. Consequently, this framework can be used to recursively test for the *goodness-of-fit* and learn its structure. For each elementary region A, the posterior stopping probability is the probability that the data distribution coincides with the base distribution conditional on A. In particular, the posterior stopping probability for the whole space Ω, $\rho(\Omega)$, measures how well the observed data fit the base overall. The posterior values of the other partitioning and pseudocount variables reflect where and how the data distribution differs from the base.

4 Coupling optional Polya trees and two sample comparison

In this section we consider the case when two i.i.d. samples are observed and one is interested in testing and characterizing the potential difference between the underlying distributions. From now on, we let Q_1 and Q_2, with densities q_1 and q_2, be the two distributions from which the two samples have come from.

4.1 Coupling optional Polya trees

A conceptually simple way to compare Q_1 and Q_2 is to proceed in two steps—first estimate the two distributions *separately*, and then use some distance metric to quantify the difference. For example, one can place an OPT prior on each of Q_1 and Q_2 and use the posteriors to estimate the densities [18]. (Other density estimators can also be used for this purpose.) With the density estimates available, one can then compute standard distance metrics such as L_1, and in turn use this as a statistic for testing the difference. (This approach provides no easy way to characterize how the two distributions are different.)

However, this two-step method is undesirable in multidimensional, and especially high-dimensional, settings. The main reason is that reliably estimating multidimensional distributions is a very difficult problem, and in fact often a much harder problem than comparing distributions. This difficulty in turn translates into either high variability or large bias in the

distance estimates, and thus low statistical power. Using this approach, one is essentially making inference on the distributional difference *indirectly*, through the inference on a large number of parameters that characterize the two distributions *per se* but have little to do with their difference.

Following this reasoning, it is favorable to make *direct* inference on "parameters" that capture the distributional difference. But such direct inference requires (from a Bayesian perspective) that the two distributions be generated from a *joint* prior. This prior should be so designed that in the corresponding joint posterior, information regarding the distributional difference can be extracted directly. We next introduce such a prior.

Our proposed method for generating the two distributions Q_1 and Q_2 is again based on a procedure that randomly partitions the space Ω and assigns probability masses into the parts, similar to the one that defines the OPT prior. What differs from the procedure for the OPT is that we add in an extra random component—the conditional coupling of the two measures Q_1 and Q_2 within the tree nodes. We next explain this construction in detail. Starting from the whole space $A = \Omega$, we draw a random variable

$$C(A) \sim Bernoulli(\gamma(A)),$$

which we call the coupling variable. If $C(A) = 1$, then we force Q_1 and Q_2 to be coupled conditional on A—that is, $Q_1(\cdot|A) = Q_2(\cdot|A)$—and we achieve this by generating a common conditional distribution from a stanford OPT on A. That is

$$Q_1(\cdot|A) = Q_2(\cdot|A) \sim OPT_{|A}(\mathcal{R}; \rho, \boldsymbol{\lambda}^b, \boldsymbol{\alpha}^b),$$

where the "b" superscript stands for "base", and the "$|A$" notation should be understood as the restriction to A of the partition rule $\mathcal{R}$, the stopping variables ρ, the partition selector variables $\boldsymbol{\lambda}^b$, and the assignment pseudo-count variables $\boldsymbol{\alpha}^b$. (For $A = \Omega$, there is no restriction.) If $C(A) = 0$, then we draw a partition selector variable

$$J(A) \in \{1, 2, \ldots, n\} \quad \text{with } P(J(A) = j) = \lambda^j(A).$$

If $J(A) = j$, then we partition A under the jth way according to $\mathcal{R}(A)$. Then draw two independent assignment vectors

$$\begin{aligned}
\boldsymbol{\theta}_1^j(A) &= (\theta_{11}^j(A), \theta_{12}^j(A), \ldots, \theta_{1K^j(A)}^j(A)) \sim Dirichlet(\alpha_{11}^j(A), \alpha_{12}^j(A), \ldots, \alpha_{1K^j(A)}^j(A)) \\
\boldsymbol{\theta}_2^j(A) &= (\theta_{21}^j(A), \theta_{22}^j(A), \ldots, \theta_{2K^j(A)}^j(A)) \sim Dirichlet(\alpha_{21}^j(A), \alpha_{22}^j(A), \ldots, \alpha_{2K^j(A)}^j(A)),
\end{aligned}$$

and let

$$Q_1(A_i^j) = Q_1(A)\theta_{1i}^j(A) \quad \text{and} \quad Q_2(A_i^j) = Q_2(A)\theta_{2i}^j(A)$$

for each child A_i^j of A. We call $\boldsymbol{\theta}_1^j(A)$ and $\boldsymbol{\theta}_2^j(A)$ the assignment vectors for Q_1 and Q_2 (in the *uncoupled* state). Then we go down one level and repeat the entire procedure for each A_i^j, starting from the drawing of the coupling variable.

Again, the first natural question to ask is whether this procedure will actually stop and give rise to two random probability measures (Q_1, Q_2). The answer is positive under very mild conditions, and this is formalized in Theorem 2. The statement of the theorem uses the notion of "forced coupling", which is similar to the idea of "forced stopping" used in the proof of Theorem 1 and which we describe next. Let $(Q_1^{(k)}, Q_2^{(k)})$ denote the pair of random distributions arising from the above random-partitioning-and-assignment procedure with forced coupling after k-levels of recursive partitioning. That is, if after k levels of partitioning a node A is reached and the two measures are not coupled on it, then force them to couple on A and generate $Q_1^{(k)}(\cdot|A) = Q_2^{(k)}(\cdot|A)$ from $OPT_{|A}(\mathcal{R}; \rho, \boldsymbol{\lambda}^b, \boldsymbol{\alpha}^b)$. We do this for all such nodes to get $(Q_1^{(k)}, Q_2^{(k)})$.

Theorem 2. *In the random-partitioning-and-assignment procedure for generating a pair of measures described above, if $\gamma(A), \rho(A) > \delta$ for some $\delta > 0$ and all potential tree nodes A defined by the partition rule $\mathcal{R}$, then with probability 1, $(Q_1^{(k)}, Q_2^{(k)})$ converges to a pair of absolutely continuous (w.r.t. μ) random probability measures (Q_1, Q_2) in the following sense.*

$$sup_{E\in\mathcal{B}}|Q_1^{(k)}(E) - Q_1(E)| + |Q_2^{(k)}(E) - Q_2(E)| \to 0,$$

where $\mathcal{B}$ is the collection of Borel sets.

Definition 2. This pair of random probability measures (Q_1, Q_2) is said to have a coupling optional Polya tree (co-OPT) distribution with partition rule $\mathcal{R}$, coupling parameters $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, γ, and base parameters $\boldsymbol{\lambda}^b$, $\boldsymbol{\alpha}^b$, ρ, and can be written as $coOPT(\mathcal{R}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \gamma; \boldsymbol{\lambda}^b, \boldsymbol{\alpha}^b, \rho)$.

Proof of Theorem 2. We first claim that with probability 1, $Q_1^{(k)}$ and $Q_2^{(k)}$ respectively converge in total variational distance to two absolutely continuous random probability measures Q_1 and Q_2, and thus for any Borel set E,

$$\begin{aligned}&|Q_1^{(k)}(E) - Q_1(E)| + |Q_2^{(k)}(E) - Q_2(E)| \\ \leq\, &sup_{E_1\in\mathcal{B}}|Q_1^{(k)}(E_1) - Q_1(E_1)| + sup_{E_2\in\mathcal{B}}|Q_1^{(k)}(E_2) - Q_1(E_2)| \to 0, \text{ w.p.1.}\end{aligned}$$

To prove the claim, we note that the marginal procedure that generates Q_1, for instance, is simply an OPT with random local base measures that arise from standard OPT distributions. To see this, we can think of the generative procedure of Q_1 as consisting of the following two steps.

1. For each potential tree node A under $\mathcal{R}$, we draw an independent random measure Q_0^A from $OPT_{|A}(\mathcal{R}, \rho, \boldsymbol{\lambda}^b, \boldsymbol{\alpha}^b)$.

2. Generate Q_1 from the standard random-partitioning-and-random-assignment procedure for an OPT, treating $\{C(A)\}$ as the stopping variables, $\{J(A)\}$ as the partition selector variables, and $\{\boldsymbol{\theta}_1^{J(A)}(A)\}$ as the probability assignment variables, and with $\{Q_0^A\}$ being the local base measures. That is, when a node A is stopped, the conditional distribution $Q_1(\cdot|A)$ is set to be $Q_0^A(\cdot)$.

By Theorem 1 in [18], for each potential node A, with probability 1, Q_0^A is an absolutely continuous distribution. Because the collection of all potential tree nodes A under $\mathcal{R}$ is countable, with probability 1, this simultaneously holds for all Q_0^A. Therefore, with probability 1, the marginal procedure for producing Q_1 is just that for an OPT with local base measures $\{Q_0^A\}$. The same argument for proving Theorem 1 in [18] (with $\mu(\cdot|A)$ replaced by $Q_0^A(\cdot)$) shows that with probability 1, an absolutely continuous measure Q_1 exists as the limit of $Q_1^{(k)}$ in total variational distance. The same argument proves the claim for Q_2 as well. □

Similar to the OPT prior, the co-OPT distribution has large support under the L_1 metric. This is formulated in the following theorem.

Theorem 3. *Let Ω be a bounded rectangle in $\mathbb{R}^p$. Suppose that the condition of Theorem 2 holds along with the following two conditions:*

(1) Under the partition rule R, the diameters of the elementary regions uniformly decreases to 0 with their levels.

(2) The coupling probabilities $\gamma(A)$, stopping probabilities $\rho(A)$, coupling selector probabilities $\lambda^j(A)$, base selection probabilities $\lambda_j^b(A)$, as well as the assignment probabilities $\alpha_{1i}^j(A)/(\sum_l \alpha_{1l}^j(A))$, $\alpha_{2i}^j(A)/(\sum_l \alpha_{2l}^j(A))$, and $\alpha_i^{bj}(A)/(\sum_l \alpha_l^{bj}(A))$ for all i, j and all potential elementary regions are uniformly bounded away from 0 and 1.

Let $q_1 = dQ_1/d\mu$ and $q_2 = dQ_2/d\mu$, then for any two density functions f_1 and f_2, and any $\tau > 0$, we have

$$P\left(\int |q_1(x) - f_1(x)|d\mu < \tau \text{ and } \int |q_2(x) - f_2(x)|d\mu < \tau\right) > 0.$$

Proof. Because any density function on Ω can be arbitrarily approximated in L_1 by uniformly continuous ones, without loss of generality, we can assume that f_1 and f_2 are uniformly continuous. Let

$$\delta_1(\epsilon) = \sup_{|x-y|<\epsilon} |f_1(x) - f_1(y)| \quad \text{and} \quad \delta_2(\epsilon) = \sup_{|x-y|<\epsilon} |f_2(x) - f_2(y)|.$$

By uniform continuity, we have $\delta_i(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$ for $i = 1, 2$. Also, by Condition (1), for any $\epsilon > 0$, there exists an integer k, such that for all nodes of level k or deeper, their diameters are less than ϵ. Let $\Omega = \cup_{i=1}^I A_i$ be a partition of Ω that can be achieved by k steps of nested partitioning. Then because the parameters of the coOPT are all bounded away from 0 and 1, there is a positive probability that the A_i's are exactly the sets on which Q_1 and Q_2 first couple. Now let q^{A_i} be the local base measure on each of A_i, we can write

$$q_1(x) = \sum_{i=1}^I Q_1(A_i) q^{A_i}(x) \mathbf{1}_{A_i}(x) \quad \text{and} \quad q_2(x) = \sum_{i=1}^I Q_2(A_i) q^{A_i}(x) \mathbf{1}_{A_i}(x).$$

Accordingly,

$$
\begin{aligned}
&\int |q_1(x) - f_1(x)| d\mu(x) \\
&= \sum_{i=1}^{I} \int_{A_i} |Q_1(A_i) q^{A_i}(x) - f_1(x)| d\mu(x) \\
&\leq \sum_{i=1}^{I} Q_1(A_i) \int_{A_i} |q^{A_i}(x) - 1/\mu(A_i)| d\mu(x) + \sum_{i=1}^{I} \int_{A_i} |Q_1(A_i)/\mu(A_i) - f_1(x)| d\mu(x) \\
&\leq \sum_{i=1}^{I} \int_{A_i} |q^{A_i}(x) - 1/\mu(A_i)| d\mu(x) + \sum_{i=1}^{I} |Q_1(A_i) - f_1^i \, \mu(A_i)| + \sum_{i=1}^{I} \int_{A_i} |f_1^i - f_1(x)| d\mu(x)
\end{aligned}
$$

where $f_1^j := \int_{A_i} f_1(x) d\mu(x)/\mu(A_i)$. By the exact same calculation we have

$$
\begin{aligned}
&\int |q_2(x) - f_2(x)| d\mu(x) \\
&\leq \sum_{i=1}^{I} \int_{A_i} |q^{A_i}(x) - 1/\mu(A_i)| d\mu(x) + \sum_{i=1}^{I} |Q_2(A_i) - f_2^i \, \mu(A_i)| + \sum_{i=1}^{I} \int_{A_i} |f_2^i - f_2(x)| d\mu(x)
\end{aligned}
$$

where $f_2^j := \int_{A_i} f_2(x) d\mu(x)/\mu(A_i)$. By the choice of A_i, we have that $\int_{A_i} |f_1^i - f_1(x)| d\mu(x) \leq \delta_1(\epsilon)\mu(A_i)$ and $\int_{A_i} |f_2^i - f_2(x)| d\mu(x) \leq \delta_2(\epsilon)\mu(A_i)$. Thus,

$$
\sum_{i=1}^{I} \int_{A_i} |f_1^i - f_1(x)| d\mu(x) \leq \delta_1(\epsilon)\mu(\Omega) \quad \text{and} \quad \sum_{i=1}^{I} \int_{A_i} |f_2^i - f_2(x)| d\mu(x) \leq \delta_2(\epsilon)\mu(\Omega).
$$

So by choosing ϵ small enough, we can have

$$
\max\{\delta_1(\epsilon), \delta_2(\epsilon)\}\mu(\Omega) < \tau/3.
$$

Next, because all the coupling parameters of the coOPT prior are uniformly bounded away from 0 and 1, (conditional on the coupling partition) with positive probability, we have

$$
|Q_1(A_i) - f_1^i \, \mu(A_i)| < \frac{\tau}{3\mu(\Omega)} \quad \text{and} \quad |Q_2(A_i) - f_2^i \, \mu(A_i)| < \frac{\tau}{3\mu(\Omega)}
$$

for all $i = 1, 2, \ldots, I$. Similarly, because all the base parameters are also uniformly bounded away from 0 and 1, by Theorem 2 in [18], (conditional on the coupling partition and probability assignments,) with positive probability we have

$$
\int_{A_i} |q^{A_i}(x) - 1/\mu(A_i)| d\mu(x) < \frac{\tau}{3 \cdot 2^i}
$$

for all $i = 1, 2, \ldots, I$. Placing the three pieces together, we have positive probability for $\int |q_1(x) - f_1(x)| d\mu < \tau$ and $\int |q_2(x) - f_2(x)| d\mu < \tau$ to hold simultaneously. □

4.2 Bayesian inference on the two sample problem using co-OPT prior

We next show that the co-OPT prior is conjugate and introduce the recipe for making inference on the two sample problem using this prior. Suppose (Q_1, Q_2) is distributed as $coOPT(\mathcal{R}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \gamma; \boldsymbol{\lambda}^b, \boldsymbol{\alpha}^b, \rho)$, and we observe two i.i.d. samples $\boldsymbol{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})$ and $\boldsymbol{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2})$ from Q_1 and Q_2 respectively. For a node A reached during the random partitioning steps in the generative procedure of (Q_1, Q_2), let $\boldsymbol{x}_1(A) = \{x_{11}, x_{12}, \ldots, x_{1n_1}\} \cap A$ and $\boldsymbol{x}_2(A) = \{x_{21}, x_{22}, \ldots, x_{2n_2}\} \cap A$ be the observations from the two samples in A, and let $n_1(A) = |\boldsymbol{x}_1(A)|$ and $n_2(A) = |\boldsymbol{x}_2(A)|$ be the sample sizes in A. As before, we let q_1 and q_2 denote the densities of the two distributions and let q_0^A denote the density of the random local base measure Q_0^A.

The likelihood of $\boldsymbol{x}_1(A)$ on A under $q_1(\cdot|A)$ and that for $\boldsymbol{x}_2(A)$ under $q_2(\cdot|A)$ are

$$\begin{cases} q_1(\boldsymbol{x}_1|A) = Cq_0^A(\boldsymbol{x}_1) + (1-C)\prod_{i=1}^{K^J}(\theta_{1i}^J)^{n_1(A_i^J)} q_1(\boldsymbol{x}_1|A_i^J) \\ \\ q_2(\boldsymbol{x}_2|A) = Cq_0^A(\boldsymbol{x}_2) + (1-C)\prod_{i=1}^{K^J}(\theta_{2i}^J)^{n_2(A_i^J)} q_2(\boldsymbol{x}_2|A_i^J) \end{cases} \tag{4.1}$$

where we have again suppressed the "(A)" notation for $C(A)$, $J(A)$, $K(A)^{J(A)}$, $\theta_{1i}^{J(A)}(A)$, $\theta_{2i}^{J(A)}(A)$, $\boldsymbol{x}_1(A)$ and $\boldsymbol{x}_2(A)$. The joint likelihood of observing $\boldsymbol{x}_1(A)$ and $\boldsymbol{x}_2(A)$ conditional on A is

$$q_1(\boldsymbol{x}_1|A)q_2(\boldsymbol{x}_2|A) = Cq_0^A(\boldsymbol{x}_1, \boldsymbol{x}_2) + (1-C)\prod_{i=1}^{K^J}(\theta_{1i}^J)^{n_1(A_i^J)}(\theta_{2i}^J)^{n_2(A_i^J)} q_1(\boldsymbol{x}_1|A_i^J)q_2(\boldsymbol{x}_2|A_i^J), \tag{4.2}$$

where $q_0^A(\boldsymbol{x}_1, \boldsymbol{x}_2) = q_0^A(\boldsymbol{x}_1)q_0^A(\boldsymbol{x}_2)$ is the standard OPT likelihood for the combined sample $\boldsymbol{x}(A) = (\boldsymbol{x}_1(A), \boldsymbol{x}_2(A))$ on A given by (2.1). Integrating out q_0^A, C, J, $\boldsymbol{\theta}_1^J$ and $\boldsymbol{\theta}_2^J$ from (4.2), we get the conditional marginal likelihood

$$P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = \gamma P_0(\boldsymbol{x}_1, \boldsymbol{x}_2|A) + (1-\gamma)\sum_{j=1}^{M}\lambda_j \frac{D(\boldsymbol{n}_1^j + \boldsymbol{\alpha}_1^j)D(\boldsymbol{n}_2^j + \boldsymbol{\alpha}_2^j)}{D(\boldsymbol{\alpha}_1^j)D(\boldsymbol{\alpha}_2^j)} \prod_{i=1}^{K^j} P(\boldsymbol{x}_1, \boldsymbol{x}_2|A_i^j), \tag{4.3}$$

where $\boldsymbol{n}_h^j = (n_h(A_1^j), n_h(A_2^j), \ldots, n_h(A_{K^j}^j)$ and $\boldsymbol{\alpha}_h^j = (\alpha_{h1}^j(A), \alpha_{h2}^j(A), \ldots, \alpha_{hK^j}^j(A))$ for $h = 1, 2$, and $P_0(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$ is the conditional marginal likelihood of the combined sample under a standard OPT as given by (2.2). Equation (4.3) provides a recursive recipe for computing the marginal likelihood term $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$ for each potential tree node A. (Of course, for this recipe to be of use, one must also specify the terminal conditions for the recursion. We will discuss ways to specify such conditions in the next subsection.)

From (4.3) one can tell that the posterior distribution of (Q_1, Q_2) is still a co-OPT distribution through the following reasoning. The first term on the RHS of (4.3), $\gamma P_0(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$, is the probability (conditional on A being a node reached in the partitioning) of the event

$$\{Q_1 \text{ and } Q_2 \text{ get coupled on } A, \text{ observe } \boldsymbol{x}_1(A) \text{ and } \boldsymbol{x}_2(A)\}.$$

The second term, $(1-\gamma)\sum_{j=1}^{M}\lambda_j\frac{D(\boldsymbol{n}_1^j+\boldsymbol{\alpha}_1^j)D(\boldsymbol{n}_2^j+\boldsymbol{\alpha}_2^j)}{D(\boldsymbol{\alpha}_1^j)D(\boldsymbol{\alpha}_2^j)}\prod_{i=1}^{K^j}P(\boldsymbol{x}_1,\boldsymbol{x}_2|A)$, is the probability of

$$\{Q_1 \text{ and } Q_2 \text{ are not coupled on } A, \text{ observe } \boldsymbol{x}_1(A) \text{ and } \boldsymbol{x}_2(A)\}.$$

Each summand, $\lambda_j\frac{D(\boldsymbol{n}_1^j+\boldsymbol{\alpha}_1^j)D(\boldsymbol{n}_2^j+\boldsymbol{\alpha}_2^j)}{D(\boldsymbol{\alpha}_1^j)D(\boldsymbol{\alpha}_2^j)}\prod_{i=1}^{K^j}P(\boldsymbol{x}_1,\boldsymbol{x}_2|A)$, is the probability (given that $C(A)=0$) of

$$\{\text{divide } A \text{ in the } j\text{th way, observe } \boldsymbol{x}_1(A) \text{ and } \boldsymbol{x}_2(A)\}.$$

Finally, given that $C(A)=0$ and $J(A)=j$, the posterior distribution for $\boldsymbol{\theta}_1^j$ and $\boldsymbol{\theta}_2^j$ are Dirichlet$(\boldsymbol{n}_1^j+\boldsymbol{\alpha}_1^j)$ and Dirichlet$(\boldsymbol{n}_2^j+\boldsymbol{\alpha}_2^j)$, respectively. This reasoning, together with Theorem 3 in [18], shows that the co-OPT prior is conjugate, and simple applications of Bayes' Theorem provide the formulae of the parameter values for the posterior. The results are summarized in the next theorem.

Theorem 4. *Suppose $\boldsymbol{x}_1=(x_{11},x_{12},\ldots,x_{1n_1})$ and $\boldsymbol{x}_2=(x_{21},x_{22},\ldots,x_{1n_2})$ are two independent i.i.d. samples from Q_1 and Q_2. Let (Q_1,Q_2) have a coOPT$(\mathcal{R};\boldsymbol{\lambda},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\gamma;\boldsymbol{\lambda}^b,\boldsymbol{\alpha}^b,\rho)$ prior that satisfies the conditions in Theorem 2. Then the posterior distribution of (Q_1,Q_2) is still a coupling optional Polya tree with the following parameters.*

1. *Coupling probabilities:*

$$\gamma(A|\boldsymbol{x}_1,\boldsymbol{x}_2)=\gamma(A)P_0(\boldsymbol{x}_1,\boldsymbol{x}_2|A)/P(\boldsymbol{x}_1,\boldsymbol{x}_2|A).$$

2. *Partition selection probabilities:*

$$\lambda_j(A|\boldsymbol{x}_1,\boldsymbol{x}_2)\propto\lambda_j(A)\frac{D(\boldsymbol{n}_1^j+\boldsymbol{\alpha}_1^j)D(\boldsymbol{n}_2^j+\boldsymbol{\alpha}_2^j)}{D(\boldsymbol{\alpha}_1^j)D(\boldsymbol{\alpha}_2^j)}\prod_{i=1}^{K^j}P(\boldsymbol{x}_1,\boldsymbol{x}_2|A),\quad j=1,2,\ldots,M(A).$$

3. *Probability assignment pseudo-counts:*

$$\alpha_{1i}^j(A|\boldsymbol{x}_1,\boldsymbol{x}_2)=n_1(A_i^j)+\alpha_{1i}^j(A)\quad \textit{and}\quad \alpha_{2i}^j(A|\boldsymbol{x}_1,\boldsymbol{x}_2)=n_2(A_i^j)+\alpha_{2i}^j(A),$$

   for $j=1,2,\ldots,M(A)$ and $i=1,2,\ldots,K^j(A)$.

4. *Base stopping probabilities:*

$$\rho(A|\boldsymbol{x}_1,\boldsymbol{x}_2)=\rho(A)u(\boldsymbol{x}_1,\boldsymbol{x}_2|A)/P_0(\boldsymbol{x}_1,\boldsymbol{x}_2|A).$$

5. *Base selection probabilities:*

$$\lambda_j^b(A|\boldsymbol{x}_1,\boldsymbol{x}_2)\propto\lambda_j^b(A)\frac{D(\boldsymbol{n}_1^j+\boldsymbol{n}_2^j+\boldsymbol{\alpha}^{bj})}{D(\boldsymbol{\alpha}^{bj})}\prod_{i=1}^{K^j}P_0(\boldsymbol{x}_1,\boldsymbol{x}_2|A),\quad j=1,2,\ldots,M(A).$$

6. *Base probability assignment pseudo-counts:*

$$\alpha_i^{bj}(A|\boldsymbol{x}_1, \boldsymbol{x}_2) = n_1(A_i^j) + n_2(A_i^j) + \alpha_i^{bj}(A),$$

for $j = 1, 2, \ldots, M(A)$ *and* $i = 1, 2, \ldots, K^j(A)$.

Two remarks: (1) All of the posterior parameter values can be computed *exactly* using the above formulae, without the need of any Monte Carlo procedure. (2) The posterior coupling parameters contain the information about the difference between the two underlying distributions Q_1 and Q_2, while the posterior base parameters contain the information regarding the underlying structure of the two measures. This naturally suggests that if one is only interested in two sample comparison, one should only need the posterior distribution of the coupling variables, and not those of the base variables. This will become clear in Sections 5 and 6 where we give several numerical examples.

4.3 Terminal conditions

As mentioned earlier, we need to specify the terminal conditions for the recursion used to compute $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$. Depending on the nature of Ω and the prior specification, the recursion formula (4.3) can terminate in several ways as demonstrated in the following two examples.

Example 1 (2^p contingency table)**.** Let $\Omega = \{1, 2\} \times \{1, 2\} \times \cdots \times \{1, 2\}$. For any rectangle A in the table—a set of the form $A_1 \times A_2 \times \ldots A_p$ with A_1, A_2, …, A_p being non-empty subsets of $\{1, 2\}$—let $k_1, k_2, \ldots, k_{M(A)}$ be the "intact" dimensions of A, that is $A_{k_j} = \{1, 2\}$ for $j = 1, 2, \ldots, M(A)$. Let $\mathcal{R}$ be the diadic cutting rule that allows A to be cut into two halves on each intact dimension j. In our earlier notation,

$$\mathcal{R}(A) = \left\{ \{A_1^j, A_2^j\}_{j=1}^{M(A)} \right\},$$

where $A_1^j = A_1 \times A_2 \times \cdots \times A_{k_j-1} \times \{1\} \times A_{k_j+1} \times \cdots \times A_p$ and $A_2^j = A_1 \times A_2 \times \cdots \times A_{k_j-1} \times \{2\} \times A_{k_j+1} \times \cdots \times A_p$. Suppose two i.i.d. samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are observed. Assume that (Q_1, Q_2) has a co-OPT prior with the following prior parameter values for each rectangle A: $\lambda_j(A) = \lambda_j^b(A) = \frac{1}{M(A)}$, $\alpha_i^j(A) = \alpha_i^{bj}(A) \equiv \frac{1}{2}$ for $i = 1, 2$ and $j = 1, 2, \ldots, M(A)$, and finally $\gamma(A) \equiv \gamma_0$, $\rho(A) \equiv \rho_0$, where γ_0 and ρ_0 are constants in $(0, 1)$.

In this example, there are three types of terminal nodes for $P_0(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$ and they are given in Example 3 of [18]. By a similar reasoning, there are also three types of terminal nodes for $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$.

1. A contains no data point from either sample. In this case, $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 1$.

2. A is an atom, that is a single cell, containing any number observations. In this case, $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 1$.

3. A contains a single observation (from either sample). In this case, $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 2^{-M(A)} = 1/\mu(A)$. To see this, first we let

$$t_{M(A)} = P(\boldsymbol{x}_1, \boldsymbol{x}_2|A).$$

By Example 3 in [18], we have

$$P_0(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 2^{-M(A)}.$$

Hence we have

$$\begin{aligned} t_{M(A)} &= \gamma_0 2^{-M(A)} + (1-\gamma_0)\left(\frac{1}{M(A)}\sum_{j=1}^{M}\frac{B\left(\frac{3}{2},\frac{1}{2}\right)}{B\left(\frac{1}{2},\frac{1}{2}\right)}\right)\cdot t_{M(A)-1} \\ &= \gamma_0 2^{-M(A)} + (1-\gamma_0)\tfrac{1}{2}t_{M(A)-1} \\ &= \gamma_0 2^{-M(A)}\frac{\left(1-(1-\gamma_0)^{M(A)}\right)}{1-(1-\gamma_0)} + \left(\frac{1-\gamma_0}{2}\right)^{M(A)} \\ &= 2^{-M(A)} = 1/\mu(A). \end{aligned}$$

Example 2 (Rectangle in $\mathbb{R}^p$)**.** Let $\Omega = I_1 \times I_2 \times \ldots \times I_p$ be a bounded rectangle in $\mathbb{R}^p$. Let $\mathcal{R}$ be the diadic partition rule such that for any rectangle A of the form $A_1 \times A_2 \times \ldots A_p$ with A_1, A_2, ..., A_p being non-empty subintervals of $I_1, I_2, \ldots, I_p$ respectively, A can be divided in half in any of the p dimensions. Again, let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the two samples, and let (Q_1, Q_2) have a co-OPT prior with the following parameters: $\lambda_j(A) = \lambda_j^b(A) \equiv \frac{1}{p}$, $\alpha_i^j(A) = \alpha_i^{bj}(A) \equiv \frac{1}{2}$, $\gamma(A) \equiv \gamma_0$ and $\rho(A) \equiv \rho_0$, for all A, $i = 1, 2$, and $j = 1, 2, \ldots, M(A)$.

In this case there are two types of terminal nodes for $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A)$.

1. A contains no observations. In this case, $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 1$.
2. A contains a single observation (from either sample). Then $P(\boldsymbol{x}_1, \boldsymbol{x}_2|A) = 1/\mu(A)$. We skip the derivation of this as it is similar to that used for Case 3 in Example 1.

Note that in this example we have implicitly assumed that no observations, from either sample, can be identical. With the assumption that Q_1 and Q_2 are absolutely continuous w.r.t. the Lebesgue measure, the probability for any observations to be identical is 0. However this situation can occur in real data due to rounding. This possibility can be dealt with in our following discussion on technical termination of the recursion.

Other than the "theoretical" terminal nodes given in the previous two examples, in real applications it is often desirable to set a technical lower limit on the size of the nodes to be computed in order to save computation. For instance, in the $\mathbb{R}^p$ example, one can impose that all nodes smaller than 1/1000 of the space Ω be stopped and coupled. That is to let $\gamma(A) = \rho(A) = 1$ by design for all small enough A. The appropriate cutoff threshold of the node size depends on the nature of the data, but typically there is a wide range of values that work well. For most problems such a technical constraint should hardly have any impact on the posterior parameter values for large nodes. It is worth emphasizing that for real-valued data, which are almost always discretized (due to rounding), such a constraint actually becomes useful also in preventing numerical anomalies. In such cases, a general rule of thumb is that one should always adopt a cutoff size larger than the rounding unit relative to the length of the corresponding boundary of the space.

5 Numerical examples on two sample comparison

The two most fundamental questions to investigate in a two sample comparison problem are

1. Did the two samples come from the same underlying distribution?

2. If the two came from different distributions, how are the two distributions different?

If one place a co-OPT prior on the pair of distributions, then these questions can be answered from the posterior co-OPT. In particular, the posterior coupling probability on the whole space, $\gamma(\Omega|\boldsymbol{x}_1, \boldsymbol{x}_2)$, is exactly the posterior probability that the two distributions are identical. In the presence of a small coupling probability on Ω, the coupling parameters reveal how and where in the space the two distributions differ.

We next provide three numerical examples, Examples 3 through 5, to demonstrate inference on the two sample problem using the co-OPT prior. In these examples, the posterior coupling probability of Ω serves as a statistic for testing whether the two samples have come from the same distribution, which we will refer to as the co-OPT statistic. In Examples 3 and 4 we compare our method to other existing approaches, and in Example 5 we show how the posterior values of the coupling variables can be used to learn the underlying structure of the discrepancy between the two samples.

In Examples 3 and 4, whenever the underlying distributions have unbounded support, we simply use the range of the data points in each dimension to define the rectangle Ω. Also, in these three examples we use 1/1000 as the size cutoff for "technical" terminal nodes as discussed in the previous section.

Example 3 (Two sample problem in $\mathbb{R}$)**.** We simulate the control and case samples under the following three scenarios.

1. Locational shift: Sample 1 $\sim$ Beta(4,6) and Sample 2 $\sim$ 0.2 + Beta(4,6) with sample sizes $n_1 = n_2 = 20$.

2. Local structure: Sample 1 $\sim$ Uniform[0,1] and Sample 2 $\sim$ 0.5 Beta(20,10) + 0.5 Beta(10,20) with $n_1 = n_2 = 30$.

3. Dispersion difference: Sample 1 $\sim$ N(0,0.5) and Sample 2 $\sim$ N(0,1) with $n_1 = n_2 = 40$.

We place a co-OPT prior on (Q_1, Q_2) as described in Example 2. (Because here there is only one dimension, there is no choice of ways to split.) We compare the ROC curves of four different statistics for testing the null hypothesis that the two samples have come from the same distribution—namely the Kolmogorov-Smirnov (KS) statistic, Cramer-von-Mises (CvM) statistic, Cramer-test statistic [2], and our co-OPT statistic. The results are presented in Figure Figure 1. Our co-OPT statistic behaves worse than the other three tests under the first scenario when there is a simple locational shift, better than CvM and Cramer tests and comparable to K-S for the second scenario, slightly worse than the Cramer test but better than the KS and CvM tests under the last scenario.

Example 4 (Two sample problem in $\mathbb{R}^2$)**.** Again we simulate two samples under three scenarios.

1. Locational shift:
   Sample 1 $\sim BN\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2^2 & 0\\0 & 2^2\end{pmatrix}\right)$ and Sample 2 $\sim BN\left(\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}2^2 & 0\\0 & 2^2\end{pmatrix}\right)$.
2. Dispersion difference:
   Sample 1 $\sim BN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$ and Sample 2 $\sim BN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.5^2 & 0\\0 & 0.5^2\end{pmatrix}\right)$.
3. Local structure:
   Sample 1 $\sim BN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5^2\\0.5^2 & 1\end{pmatrix}\right)$, and
   Sample 2 $\sim 0.5 \times BN\left(\begin{pmatrix}0.5\\0.5\end{pmatrix}, \begin{pmatrix}0.4^2 & 0\\0 & 0.4^2\end{pmatrix}\right) + 0.5 \times BN\left(\begin{pmatrix}-0.5\\-0.5\end{pmatrix}, \begin{pmatrix}0.4^2 & 0\\0 & 0.4^2\end{pmatrix}\right)$.

The sample sizes are $n_1 = n_2 = 50$ for each scenario. We place a co-OPT prior on (Q_1, Q_2) as described in Example 2. We compare the ROC curve of our co-OPT statistic to that of the Cramer test [2]. The results are presented in Figure Figure 2. Our test is again less powerful than the Cramer test for locational shift, but more powerful under the other two scenarios.

Our next example deals with retrospectively sampled data on a high-dimensional contingency table. In this example, we not only demonstrate the power of our method to test for two sample difference, but also show that the posterior co-OPT distribution can help learn the underlying structure of the difference.

Example 5 (Retrospectively sampled data on a 2^{15} contingency table)**.** Suppose there are 15 binary predictors $X_1, X_2, \ldots, X_{15}$, and there is a binary response variable Y, e.g. disease status, whose distribution is

$$Y \sim \begin{cases} Bernoulli(0.3) & \text{if } X_3 = 1 \text{ and } X_7 = 1 \\ Bernoulli(0.3) & \text{if } X_7 = 0 \text{ and } X_{10} = 0 \\ Bernoulli(0.1) & \text{otherwise.} \end{cases}$$

We simulate populations for joint observations of X_i's and Y of size 20,000 under two scenarios

1. $X_1, X_2, \ldots, X_{15} \sim_{i.i.d.}$ Bernoulli(0.5)
2. $X_1, X_2, \ldots X_8$ as a Markov Chain with $X_1 \sim$ Bernoulli(0.5), and $P(X_t = X_{t-1}|X_{t-1}) = 0.7$, while $X_9, X_{10}, \ldots X_{15} \sim_{i.i.d}$ Bernoulli(0.5) and are independent of $X_1, \ldots, X_8$.

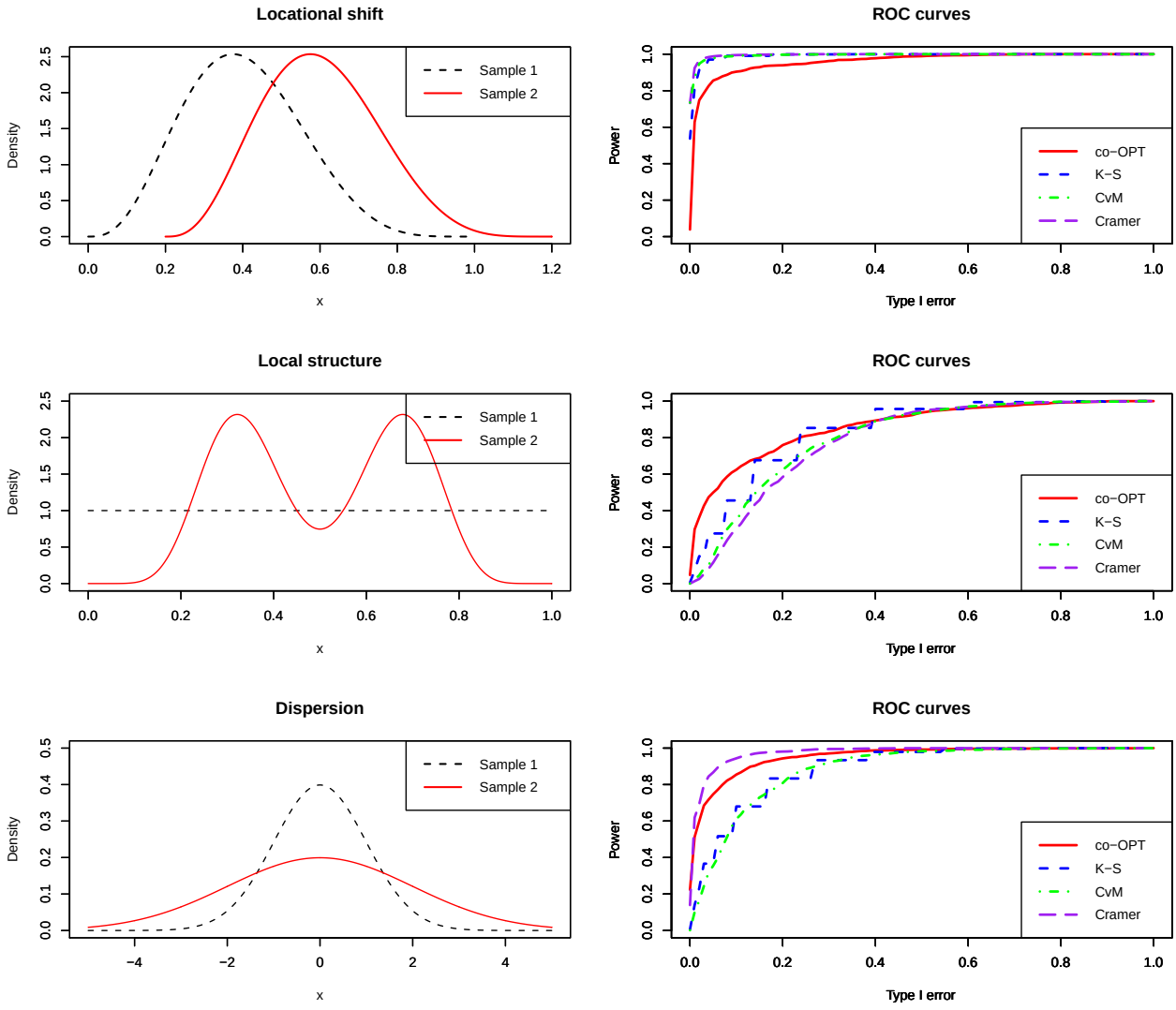

Figure 1: Two simulated samples on $\mathbb{R}$ under three scenarios (rows). (1) Locational shift: Sample 1 from $\sim$ Beta(4,6), and Sample 2 from $\sim$ 0.2 + Beta(4,6). Sample sizes: $n_1 = n_2 = 20$. (2) Local structure: Sample 1 from $\sim$ Uniform[0,1], and Sample 2 from $\sim 0.5Beta(20, 10) + 0.5Beta(10, 20)$. Sample sizes: $n_1 = n_2 = 30$. (3) Dispersion difference: Sample 1 from $N(0, 0.5)$, and Sample 2 from $N(0, 1)$. Sample sizes: $n_1 = n_2 = 40$. Left panel: Density functions for the two samples. Right panel: ROC curves for four different tests.

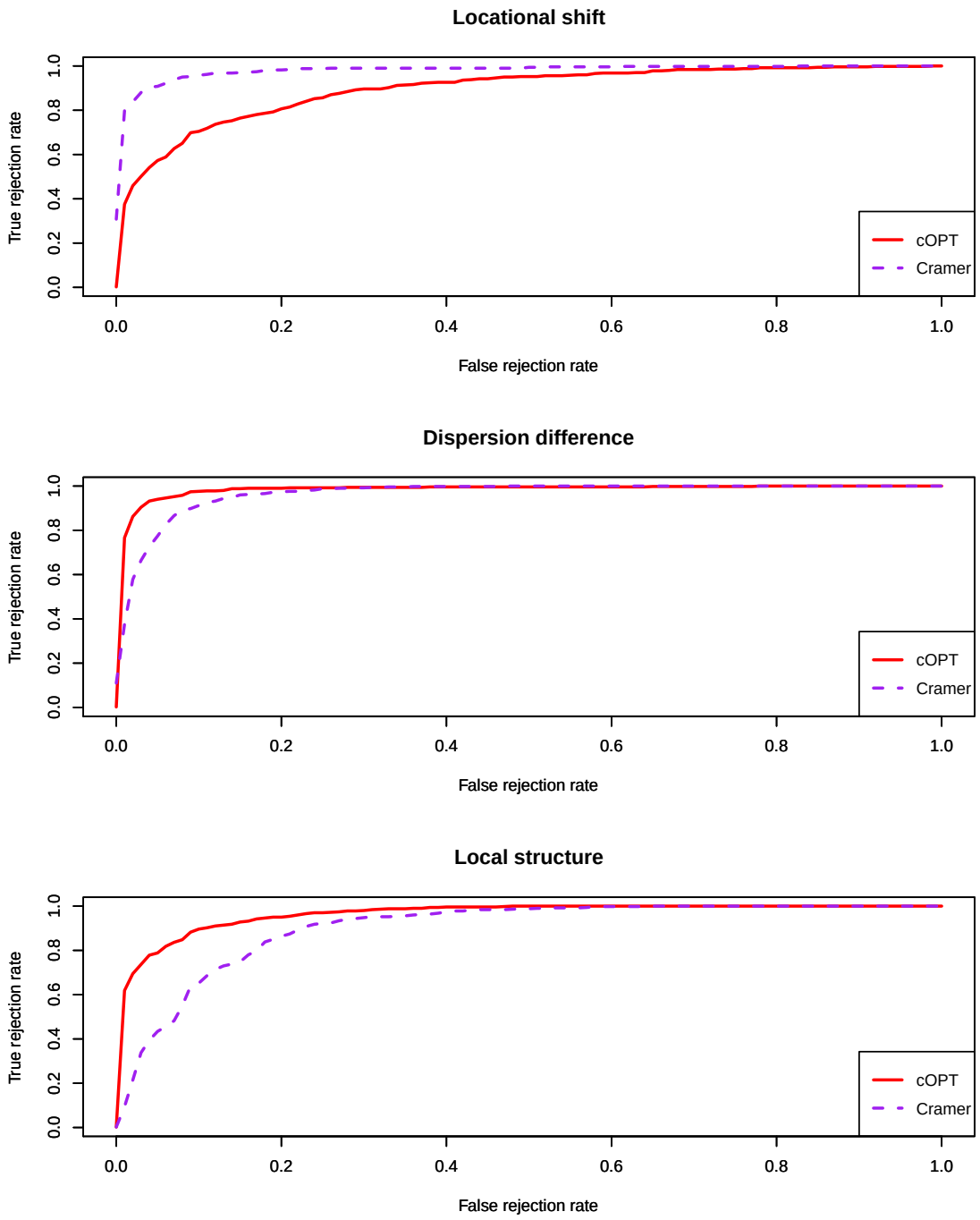

Figure 2: ROC curves for two simulated samples on $\mathbb{R}^2$ under the three scenarios given in Example 4.

For each scenario, we retrospectively sample controls (Y=0) and cases (Y=1). Our interest is in (1) the power of our method in detecting the difference in the joint distribution of the predictor variables between the two samples, and (2) whether the method can recover the "interactive" structure among the three predictors X_3, X_7 and X_{10}.

We place a co-OPT prior on (Q_1, Q_2) as described in Example 1. Figure Figure 3 presents the ROC curves of co-OPT statistic for three different sets of sample sizes under Scenarios 1 and 2. As we can see, the correlation among the markers under the second scenario does cost some of the power, but even then, just 500 data points in each sample are sufficient for achieving high power. In addition to detecting the difference, the posterior co-OPT captures the underlying structure of the difference. We find that with about 500 data points in each sample for Scenario 1 and about 3500 data points in each sample for Scenario 2, the underlying structure can be accurately recovered using the hierarchical *maximum a posteriori* (hMAP) tree topology, which is a top-down stepwise posterior maximum likelihood tree. (The construction of the hMAP tree as well as the motivation to choose it over the MAP tree is discussed in detail in Section 4.2 of [18].) As one would expect, the correlation between the predictor variables makes it much harder to recover the exact interactive relation. A typical hMAP tree structure for the simulated populations with these sample sizes is given in Figure Figure 4. We note that a sample from the posterior distribution of the tree structure can be more informative than the hMAP tree, especially when the sample sizes are not large enough. We use the hMAP here as a demonstration for its ease of visualization.

6 Inference on distributional distances between two samples

In some situations, one may be interested in a distance measure for the two sample distributions. For example, if we let $d(Q_1, Q_2)$ denote the distance between the two sample distributions under some metric d, one may want to compute quantities such as $P(d(Q_1, Q_2) > T)$ where T is some constant. This can be achieved if one knows the posterior distribution of $d(Q_1, Q_2)$ or can sample from it. We next show that if (Q_1, Q_2) arises from a co-OPT distribution, then for some common metrics, in particular L_1 and Hellinger distances, it is very convenient to sample from the distribution of $d(Q_1, Q_2)$.

As before, let Q_1 and Q_2 (with densities q_1 and q_2 respectively) be the two distributions of interest. Suppose (Q_1, Q_2) have a co-OPT distribution, and so they can be thought of as being generated from the random-partitioning-and-assignment procedure introduced in the previous section through the drawing of the variables C, J, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, C^b, J^b and $\boldsymbol{\theta}^b$. Then we have the following key proposition.

Proposition 5. *Suppose (Q_1, Q_2) has a co-OPT distribution satisfying the conditions given in Theorem 2. Let $\mathcal{A}(C, J)$ denote the (random) collection of all nodes on which Q_1 and Q_2 first couple. (The notation indicates that it depends on the coupling variables C and J.)*

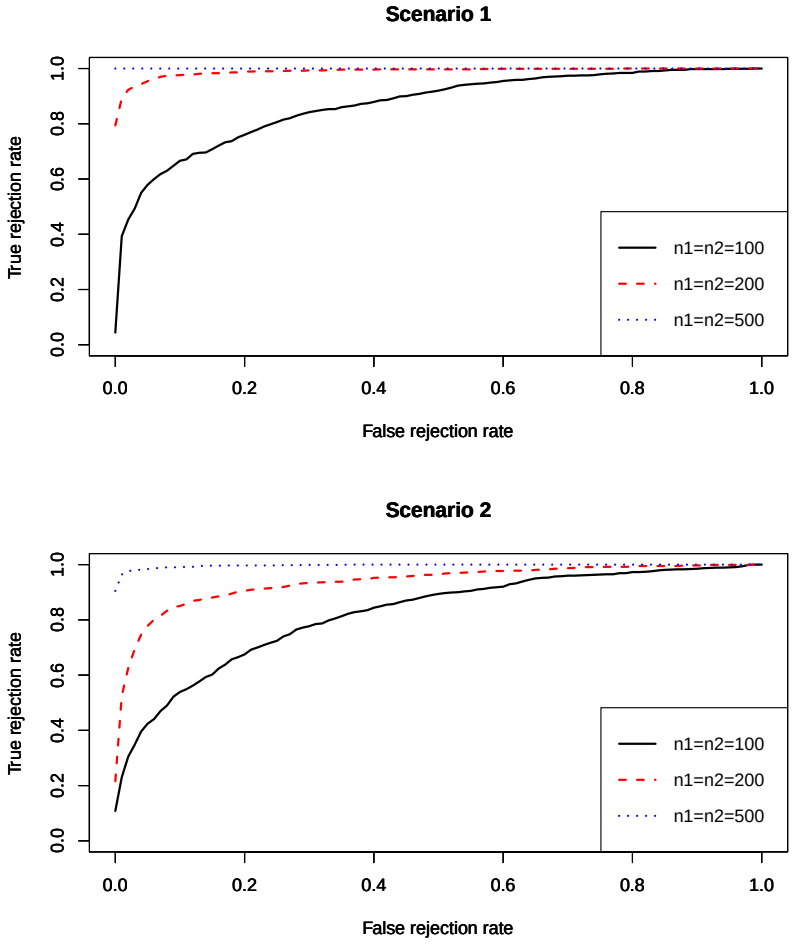

Figure 3: ROC curves of the co-OPT statistic for Example 5.

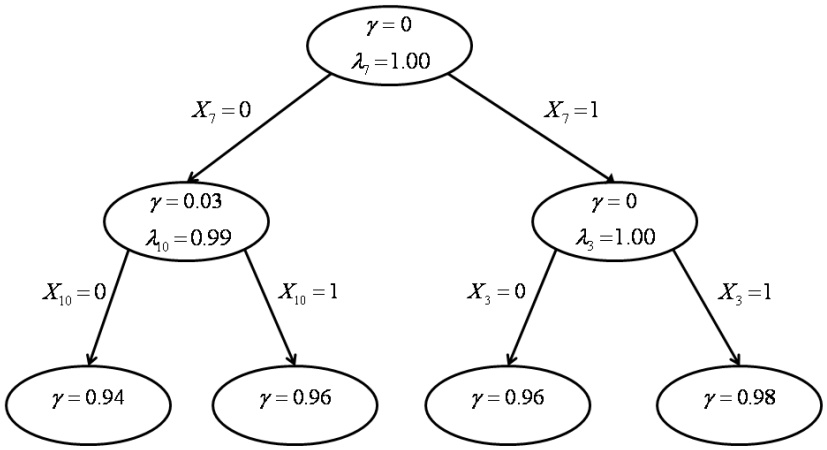

Figure 4: A typical hMAP coupling tree that recovers the underlying interactive structure.

Also, let d_{L_1} be the L_1 distance, and d_{H^2} the (squared) Hellinger distance. Then

$$d_{L_1}(Q_1, Q_2) = \sum_{A \in \mathcal{A}(C,J)} |Q_1(A) - Q_2(A)|$$
$$d_{H^2}(Q_1, Q_2) = \sum_{A \in \mathcal{A}(C,J)} (\sqrt{Q_1(A)} - \sqrt{Q_2(A)})^2.$$

Proof. We prove the result only for d_{L_1} as the proof for d_{H^2} is very similar. (All following equalities and statements hold with probability 1.)

$$\begin{aligned} d_{L_1}(Q_1, Q_2) &= \int_\Omega |q_1(x) - q_2(x)| \mu(dx) \\ &= \sum_{A \in \mathcal{A}(C,J)} \int_A |q_1(x) - q_2(x)| \mu(dx) + \int_{\Omega \setminus \cup \mathcal{A}(C,J)} |q_1(x) - q_2(x)| \mu(dx). \end{aligned}$$

But for each $A \in \mathcal{A}(C, J)$, due to coupling we have $q_1(\cdot|A) = q_2(\cdot|A)$, and so

$$\begin{aligned} \int_A |q_1(x) - q_2(x)| \mu(dx) &= \int_A |Q_1(A) - Q_2(A)| q_1(x|A) \mu(dx) \\ &= |Q_1(A) - Q_2(A)|. \end{aligned}$$

On the other hand, $Q_1(\Omega \setminus \cup \mathcal{A}(C, J)) = Q_2(\Omega \setminus \cup \mathcal{A}(C, J)) = \mu(\Omega \setminus \cup \mathcal{A}(C, J)) = 0$ w.p.1. (See proof of Theorem 1 in [18].) Therefore,

$$d_{L_1} = \sum_{A \in \mathcal{A}(C,J)} |Q_1(A) - Q_2(A)|.$$

□

This proposition provides a recipe for drawing samples from the distributions of $d_{L_1}(Q_1, Q_2)$ and $d_{H^2}(Q_1, Q_2)$. More specifically, first draw the coupling variables C, J, $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$. Then use C and J to find the collection of nodes $\mathcal{A}(C, J)$, and use $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ to compute, for each $A \in \mathcal{A}(C, J)$, the corresponding measures $Q_1(A)$ and $Q_2(A)$. Finally, compute one draw of d_{L_1} (or d_{H^2}) by summing $|Q_1(A) - Q_2(A)|$ (or $(\sqrt{Q_1(A)} - \sqrt{Q_2(A)})^2$) over all nodes in $\mathcal{A}(C, J)$.

A particularly desirable feature of this procedure for sampling L_1 and Hellinger distances is that one does not need to draw samples for the two random distributions Q_1 and Q_2 to get their distances. In fact, one only needs to draw the coupling variables, which characterize the *difference* between the two distributions, without having to draw the base variables, which characterize the fine structure of the two densities. Again, in multi-dimensional settings where estimating densities is difficult, such a procedure can produce much less variable samples for the distances.

We close this section with two more numerical examples, one in $\mathbb{R}$ and one in $\mathbb{R}^2$. In these two examples, again we use the observed range of the data in each dimension to define the space Ω. Also, we use 1/10000 as the size cutoff for technical termination.

Example 6 (Two beta distributions)**.** We simulate two samples from Beta(2,5) and Beta(20,15) under three sets of sample sizes $n_1 = n_2 =$10, 100 and 1000. We place a co-OPT prior on the two distributions with the diadic partition rule and the symmetric parameter values as specified in Example 2 with $\rho_0 = \gamma_0 = 0.5$, and compute the corresponding posterior co-OPT. Then we draw 1000 samples for each of $d_{L_1}(Q_1, Q_2)$ and $d_{H^2}(Q_1, Q_2)$ from their posterior distributions. The histograms of these samples are plotted in Figure Figure 5, where the vertical lines indicate the actual L_1 and Hellinger distances between the two distributions.

Example 7 (Bivariate normal and mixture of bivariate normal)**.** We repeat the same thing as in the previous example except now we simulate the two samples from the following distributions in $\mathbb{R}^2$.

Sample 1 $\sim BN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 4\end{pmatrix}\right)$, and

Sample 2 $\sim 0.5 \times BN\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) + 0.5 \times BN\left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$.

Again we draw 1000 posterior samples for $d_{L_1}(Q_1, Q_2)$ and for $d_{H^2}(Q_1, Q_2)$ under each set of sample sizes. The histograms of these samples are plotted in Figure Figure 6, where the vertical lines again indicate the actual L_1 and Hellinger distances between the two distributions.

7 Concluding remarks

In this work we have introduced the coupling optional Polya tree prior for Bayesian nonparametric analysis on the two sample problem. This prior jointly generates two random

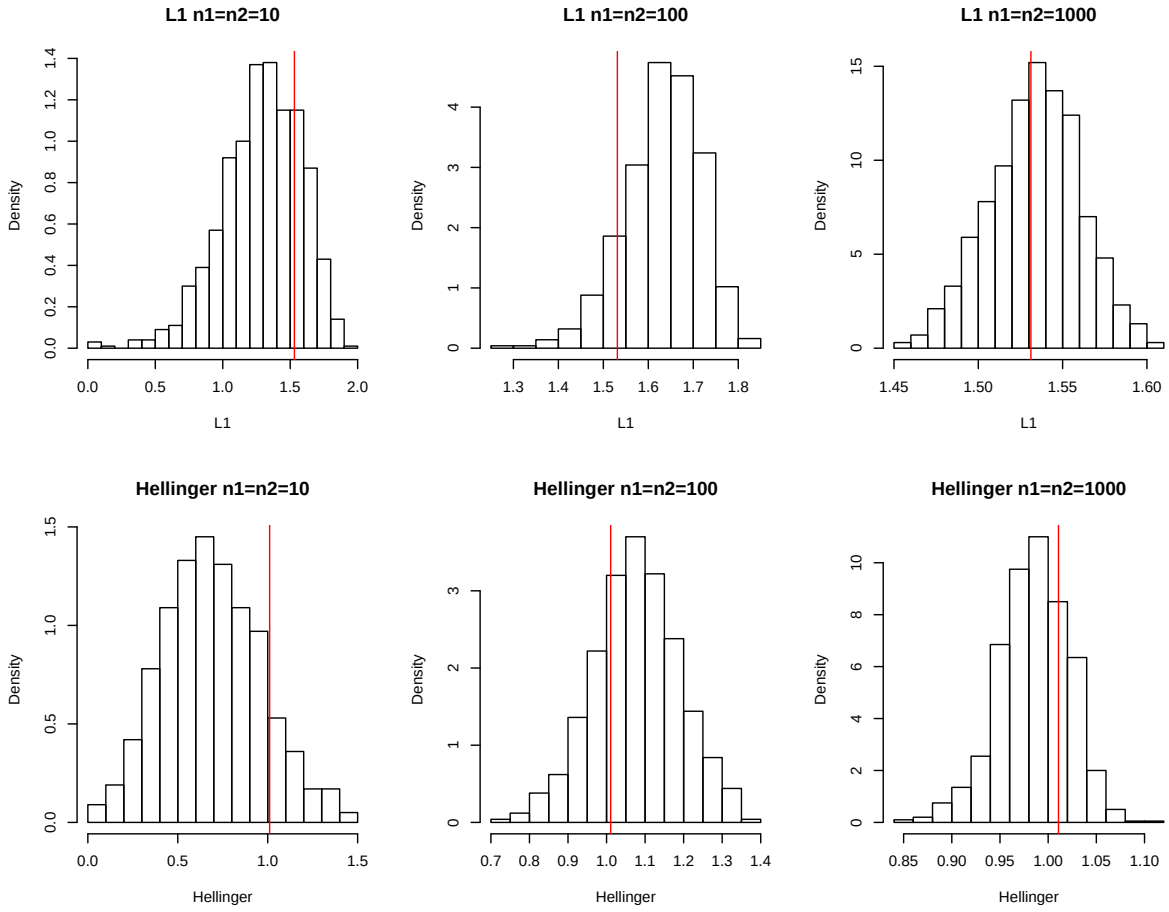

Figure 5: Histograms for posterior samples of L_1 and squared Hellinger distances for two samples from Beta(2,5) and Beta(20,15). The vertical lines indicate the actual L_1 and squared Hellinger distance between these two distributions.

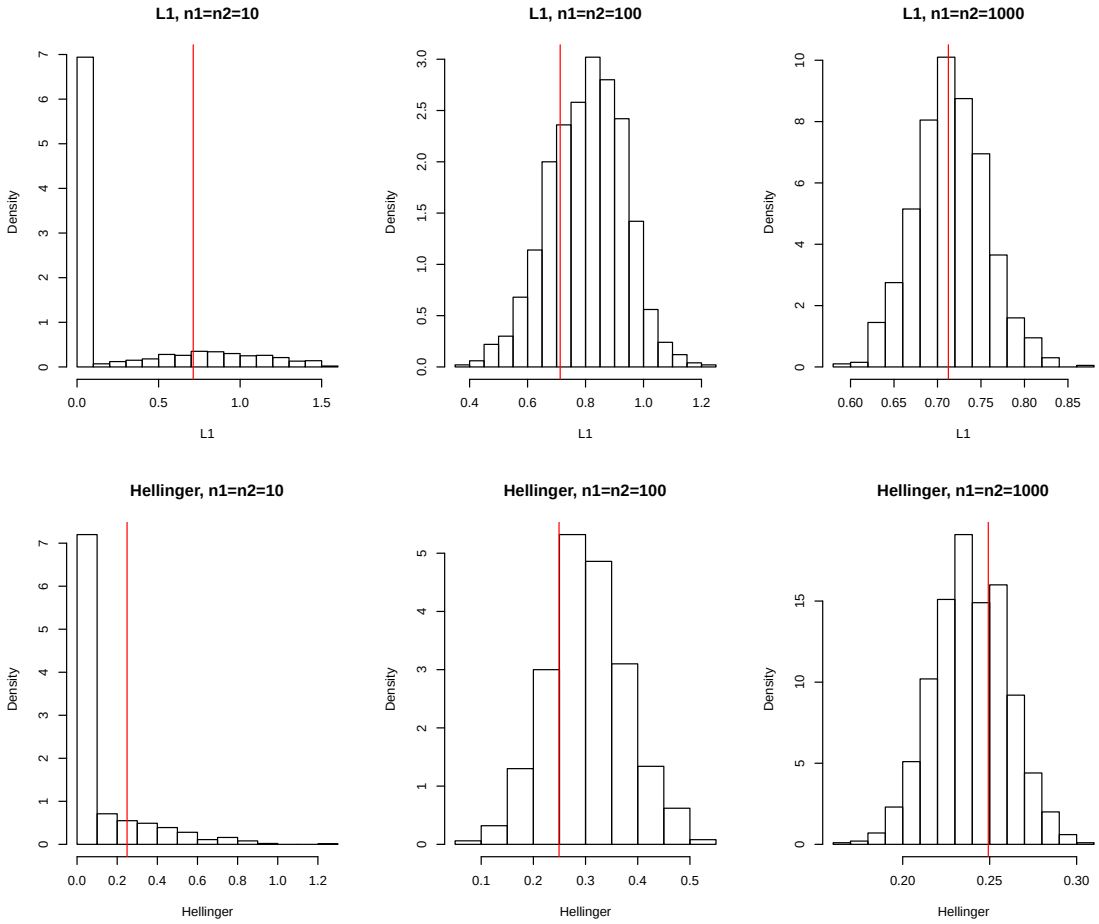

Figure 6: Histograms for posterior samples of L_1 and squared Hellinger distances for Example 7. The vertical lines indicate the actual L_1 and squared Hellinger distances for the two underlying distributions.

probability distributions that can "couple" on subsets of the state space. We have demonstrated that this construction allows both the testing and the learning of the distributional difference between the two samples. One can easily extend this prior to allow the joint generation of more than two samples. For example, if four samples are involved, then one can draw four, instead of two, independent Dirichlet vectors for probability assignment on each uncoupled node.

One interesting feature of the co-OPT prior (as well as the original OPT prior) is that the corresponding posterior can be computed "exactly" using the recursive formulation given in (4.2) without resorting to Monte Carlo sampling. However, such "exact inference" [11] based on recursions is still computationally intensive, especially in high-dimensional problems. Efficient implementation is a necessity for this method to be feasible for any non-trivial problems. More detailed discussion on the computational issues and ways to efficiently implement the inference can be found in LM's dissertation. However, even with the most efficient implementation, the exponential nature of the method dictates that approximation techniques such as k-step look-ahead as well as large-scale parallelization are needed for "ultra-high" dimensional problems, such as those on a contingency table with 100 dimensions. Current work is undergoing in this direction.

Acknowledgment

The authors want to thank Art Owen, David Siegmund, Hua Tang, and Robert Tibshirani for very helpful comments. LM is support by a Larry Yung Stanford Interdisciplinary Fellowship in Human Health. WHW is supported in part by NSF grants DMS-0505732 and DMS-0906044. Much of the computation is done on a system supported by an NSF SCREMS Grant DMS-0821823 to the Stanford Department of Statistics.

References

[1] Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Stat.* **33**, 3, 1148–1159.

[2] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivar. Anal.* **88**, 1, 190–206.

[3] Basu, S. and Chib, S. (2003). Marginal likelihood and bayes factors for dirichlet process mixture models. *Journal of the American Statistical Association* **98**, 461, 224–235.

[4] Berger, J. O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association* **96**, 453, 174–184.

[5] Carota, C. and Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In *Bayesian statistics 5*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Eds. Oxford University Press, 507–511.

[6] Chakravarti, I. M., Laha, R. G., and Roy, J. (1967). *Handbook of Methods of Applied Statistics*. Vol. **I**. John Wiley and Sons, USE.

[7] Florens, J. P., Richard, J. F., and Rolin, J. M. (1996). Bayesian encompassing specification tests of a parametric model against a non parametric alternative. Technical Report 96.08, Université Catholique de Louvain, Institut de statistique.

[8] Griffin, J. and Steel, M. (2006). Order-based dependent dirichlet processes. *Journal of the American Statistical Association 101*, 179–194.

[9] Hanson, T. E. (2006). Inference for mixtures of finite pólya tree models. *Journal of the American Statistical Association 101*, 1548–1565.

[10] Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2009). Two-sample Bayesian nonparametric hypothesis testing. http://arxiv.org/abs/0910.5060.

[11] Hutter, M. (2009). Exact non-parametric Bayesian inference on infinite trees. Tech. Rep. 0903.5342, ARXIV. http://arxiv.org/abs/0903.5342.

[12] Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer, New York.

[13] MacEachern, S. (1999). Dependent dirichlet processes. In *Proceedings of the section on Bayesian Statistical Science*.

[14] McVinish, R., Rousseau, J., and Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandinavian Journal of Statistics* **36**, 2, 337–354.

[15] 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* **66**, 3, 735–749.

[16] 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.

[17] Tokdar, S. T., Chakrabarti, A., and Ghosh, J. K. (2010). Bayesian nonparametric goodness of fit tests. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, M.-H. Chen, D. K. Dey, P. Mueller, D. Sun, and K. Ye, Eds.

[18] Wong, W. H. and Ma, L. (2010). Optional Pólya tree and Bayesian inference. *Annals of Statistics* **38**, 3, 1433–1459. http://projecteuclid.org/euclid.aos/1268056622.