Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **2**, Number **2**, June **2020**
doi:10.3934/fods.2020009
pp. **155–172**

# A BAYESIAN NONPARAMETRIC TEST FOR CONDITIONAL INDEPENDENCE

Onur Teymur and Sarah Filippi

Department of Mathematics
Imperial College London, UK

Abstract. This article introduces a Bayesian nonparametric method for quantifying the relative evidence in a dataset in favour of the dependence or independence of two variables conditional on a third. The approach uses Pólya tree priors on spaces of conditional probability densities, accounting for uncertainty in the form of the underlying distributions in a nonparametric way. The Bayesian perspective provides an inherently symmetric probability measure of conditional dependence or independence, a feature particularly advantageous in causal discovery and not employed in existing procedures of this type.

1. **Introduction.** The random variables $X$ and $Y$ are conditionally independent given $Z$ (written $X \perp\!\!\!\perp Y \mid Z$) if and only if the following relation holds between their conditional densities, for all possible realised values $z$ of $Z$:

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) \cdot p_{Y|Z}(y|z) \tag{1}$$

A common problem in the analysis of multi-variable datasets is that of assessing whether or not this relation is true for a given triple of variables. Typically, the setting is that the three densities in (1)—and the marginal density $p_Z(z)$—are all unknown *a priori*, but we have a finite set of data $W := \{(X_i, Y_i, Z_i)\ ;\ i = 1, \dots, N\}$ assumed to be drawn from the joint measure $p_{XYZ}$ induced by $(X, Y, Z)$. Notably, this type of analysis is a key component in most common approaches to causal discovery [28].

Testing for conditional independence with finite data is, however, known to be a hard problem in general. This is particularly true if the unknown densities are assumed continuous and modelled nonparametrically. In such a setting, a test for conditional independence with desirable statistical properties cannot in general be constructed [2, 34]. Nonetheless, many tests exist and are commonly used in practice, despite their various theoretical deficiencies. A classic approach is to form a test statistic from the partial correlation coefficient [8]. This vanishes if $X \perp\!\!\!\perp Y \mid Z$, but only under the strong assumptions that all variables are Gaussian and all dependences linear. Only limited extensions for non-Gaussian variables [15, 31] and for nonlinear dependences [16, 30] exist. Other approaches include combining a series of unconditional independence tests on the response variables $(X, Y)$ conditional on multiple individual values $z$ of $Z$ [25, 17]; tests based on measures of statistical distance between estimates of the conditional densities $p_{X|Z}$ and $p_{X|YZ}$, which are zero if and only if $X \perp\!\!\!\perp Y \mid Z$ [37, 38]; tests based on

2020 *Mathematics Subject Classification.* Primary: 62-08; Secondary: 62G10.
*Key words and phrases.* Bayesian nonparametrics, conditional independence testing.
Supported by EPSRC grant EP/R013519/1.

estimation of the conditional mutual information of $X$ and $Y$ given $Z$ [19, 32, 33]; permutation-type tests [4, 3] that require knowledge of or estimation of $p_{X|Z}$; and a large range of kernel-based methods [9, 41, 5, 40, 36] typically designed with the aim of dealing with high-dimensional or sparse problems more effectively.

All of the methods described in the previous paragraph are frequentist by construction, in that they derive a test statistic and construct a hypothesis test based on either a known null distribution, an asymptotic approximation to it, or by using some other strategy such as a permutation test. In the latter case, the issue is complicated by the fact that permutation tests are not easy to design in the setting of conditional independence testing with continuous $Z$ variable, an issue addressed by a range of modified non-uniform permutation tests [32, 4, 3].

Whichever specific method is used, Peters et al. [29, §7.2.1] point out one possible problem with relying on a frequentist testing procedure for causal inference, namely that "all causal discovery methods that are based on conditional independence tests draw conclusions both from dependences and independences". This reminds us that classical hypothesis testing is inherently asymmetric. Specifically, it is often necessary to detect situations in which the data are 'in favour of the null hypothesis' of conditional independence—this is how the PC algorithm [35] determines which edges to remove in the process of recovering a causal graph. However, doing so subtly abuses the classical hypothesis testing framework, in which one cannot directly compute evidence in favour of the null hypothesis.

Bayesian hypothesis testing circumvents this issue. To the best of our knowledge there is only a very limited existing literature in Bayesian testing for conditional independence—a method for the case of Gaussian random variables only, for which conditional independence is equivalent to zero partial correlation [12].[1] In this paper, we propose the first Bayesian nonparametric approach for conditional independence testing. The procedure produces a probabilistic measure of the relative evidence in a dataset for dependence or independence of two random variables $X$ and $Y$ conditionally on a third variable $Z$. The nonparametric approach permits the computation of such a probabilistic measure without assuming a known form for the underlying conditional distribution $p_{XY|Z}$. Following Filippi & Holmes [7], who construct a Bayesian nonparametric test for (unconditional) independence, we use Pólya tree priors to model the unknown data-generating distributions.

1.1. **Bayesian nonparametric hypothesis testing.** Recall that we have a dataset $W := \{(X_i, Y_i, Z_i) : i = 1, \dots, N\}$ and wish to compare two competing hypotheses $H_0$ and $H_1$, with $H_0$ the hypothesis of conditional independence and $H_1$ the contrary.

$$\begin{aligned} H_0 &: X \perp\!\!\!\perp Y \mid Z \\ H_1 &: X \not\perp\!\!\!\perp Y \mid Z \end{aligned} \tag{2}$$

Our aim is to quantify the relative evidence for these hypotheses in the dataset $W$, which is naturally measured by the posterior probabilities $p(H_0|W)$ and $p(H_1|W)$.

[1] There are algorithms among those surveyed in this section that can be viewed as a 'halfway house' towards the Bayesian ideal. In [19], for instance, the authors derive a posterior distribution over the conditional mutual information between $X$ and $Y$ given $Z$, which they treat as random in a Bayesian manner. However the output of their method does not directly provide a posterior probability in favour of one of the competing hypotheses.

To evaluate these posterior probabilities, we use the Bayes Factor [18], defined as the ratio of the marginal likelihoods of two conditional data-generating models.

$$\mathrm{BF}(H_0, H_1) = \frac{p_{XY|Z}(W|H_0)}{p_{XY|Z}(W|H_1)} \tag{3}$$

With the prior probabilities of the two hypotheses denoted by $p(H_0)$ and $p(H_1)$, we can use this to derive the posterior probability of $H_1$ as

$$p(H_1|W) = \frac{1}{1 + \mathrm{BF}(H_0, H_1)p(H_0)p(H_1)^{-1}} \tag{4}$$

The ratio of marginal likelihoods on the right-hand side of (3) can be expanded by factorising the numerator. This is simply an application of the definition of conditional independence given by (1).

$$\frac{p_{XY|Z}(W|H_0)}{p_{XY|Z}(W|H_1)} = \frac{p_{X|Z}(W|H_0)p_{Y|Z}(W|H_0)}{p_{XY|Z}(W|H_1)} \tag{5}$$

In the remainder we suppress the explicit marking of the models $H_0$ and $H_1$ since the subscripts now make clear which of the three terms belongs to which model.

We now follow a Bayesian nonparametric approach to accommodate the uncertainty in the form of the three unknown conditional densities on the right-hand side of (5). For a domain $\Omega$, we denote by $\mathcal{M}(\Omega)$ the space of all probability measures on $\Omega$. Consider first the two-dimensional conditional density $p_{XY|Z}$ (corresponding to hypothesis $H_1$) with $X$, $Y$ and $Z$ all univariate real random variables. The Bayesian nonparametric approach entails placing a functional prior $\pi$ on $\mathcal{M}(\mathbb{R}^2 \times \mathbb{R})$—individual elements of which we call $q(\cdot|\cdot)$—incorporating the data $W$ through a likelihood function $\mathcal{L}$, then marginalising over $\mathcal{M}(\mathbb{R}^2 \times \mathbb{R})$ such that the *conditional marginal likelihood* is given by

$$\begin{aligned} p_{XY|Z}(W) &= \int_{\mathcal{M}(\mathbb{R}^2\times\mathbb{R})} \mathcal{L}(W; q)\, d\pi(q) \\ &= \int_{\mathcal{M}(\mathbb{R}^2\times\mathbb{R})} \prod_{i=1}^{N} q(X_i, Y_i|Z_i)\, d\pi(q)\ ; \end{aligned} \tag{6}$$

We refer the reader to the comprehensive textbook treatments in [10, 11] for further details on the basic principles of the Bayesian nonparametric approach. The same procedure is now applied to the one-dimensional conditional densities $p_{X|Z}$ and $p_{Y|Z}$ with $\pi$, $q \in \mathcal{M}(\mathbb{R} \times \mathbb{R})$ and $\mathcal{L}$ replaced by their one-dimensional analogues.

Since we wish to assume that the random variables are all continuous, we select $\pi$ to be from the Pólya tree family of priors. These priors are supported on the entire space of probability measures $\mathcal{M}(\Omega)$ [11, Thm. 3.3.6] and can be designed to ensure that individual samples $q$ are absolutely continuous with probability one. Furthermore, they have the advantage that the marginal likelihood in (6) is tractable, in contrast to other nonparametric models of continuous random variables such as the Dirichlet Process Mixture [6].

The specific model we use is a modified version of the conditional Optional Pólya tree (cond-OPT) of Ma [23], also incorporating ideas from the finite Pólya tree of Lavine [21] and the multi-dimensional Pólya tree of Paddock [27]. We review these models in the coming sections. The first constructions we explore are designed for modelling random *unconditional* density functions $q(\cdot)$; later we will see how to build upon these to model random *conditional* density functions $q(\cdot|\cdot)$.

2. **Pólya trees.** The classical (unconditional) Pólya tree (PT) [20, 26, 21] essentially defines a random probability measure over a one-dimensional domain $\Omega \subseteq \mathbb{R}$. The most familiar construction proceeds by recursive binary partitioning of $\Omega$ and at each step the assigning of probability mass to the two child sets of a set $C \subseteq \Omega$ by means of independent Beta-distributed random branching variables $\theta$. This results in a tree structure, similar to that shown in Figure 1. Constructed this way, it is helpful to think of the Pólya tree as a random histogram on $\Omega$ or, for parameter choices which result in continuous distributions almost surely, a random density function. A particle of probability mass can be thought of as cascading down the tree, with the direction it takes at each binary split determined by the random parameters $\theta$.

More precisely, let $q$ denote a random probability density[2] on $\Omega$ and $\pi$ a measure over $\mathcal{M}(\Omega)$. Consider a partitioning of $\Omega$ in two disjoint sets $C_0$ and $C_1$, define the random branching probability $\theta_0 \equiv q(C_0) \sim \text{Be}(\alpha_1, \alpha_1)$ for some $\alpha_1 > 0$. It follows that $\theta_1 \equiv q(C_1) = 1 - \theta_0$. Note that in general, the two parameters of this Beta distribution need not be the same, though this symmetrising simplification is common and we adopt it. Indeed, we take the parameters constant within each level of the tree; the subscript on $\alpha_j$ denotes this level. Continue in this fashion, with $C_0 = C_{00} \cup C_{01}$, $C_{00} \cap C_{01} = \emptyset$ and $\theta_{00} \equiv q(C_{00}|C_0) \sim \text{Be}(\alpha_2, \alpha_2)$, $\theta_{000} \equiv q(C_{000}|C_{00}) \sim \text{Be}(\alpha_3, \alpha_3)$ and so on recursively, with each independent Beta random variable $\theta_*$ determining the probability that the particle enters the set $C_*$ at the next level of the tree.

We write $\varepsilon_i$ for a (single) element of the set $\{0, 1\}$, $\varepsilon^j \equiv \varepsilon_1 \varepsilon_2 \dots \varepsilon_j$ for a length-$j$ word from the set $\{0, 1\}^j$, $\varepsilon^j 0$ and $\varepsilon^j 1$ for the appending of respectively a single 0 or 1 onto the end of $\varepsilon^j$, and $E^j$ for the set of all length-$j$ $\{0, 1\}$–words. We further write $\varepsilon^*$ for an element of the set $E^* \equiv \bigcup_{j=1}^{\infty} E^j$ of all possible $\{0, 1\}$–words of *any* finite length. The measure of a set $C_{\varepsilon_1 \varepsilon_2 \dots \varepsilon_j}$ can then be written as

$$q(C_{\varepsilon_1 \varepsilon_2 \dots \varepsilon_j}) = \prod_{i=1}^{j} q(C_{\varepsilon_1 \varepsilon_2 \dots \varepsilon_i} | C_{\varepsilon_1 \varepsilon_2 \dots \varepsilon_{i-1}}) \tag{7}$$

Taking the infinite limit of tree depth $j$, it can be shown that the set of finite unions of intervals of the form $C_{\varepsilon^*}$ generates the Borel $\sigma$-algebra on $\Omega$. With $q$ constructed in this fashion, the measure $\pi(q)$ is a Pólya tree.

Under certain conditions on the parameters $\alpha$, the Pólya tree assigns positive probability to the Kullback–Leibler neighbourhood of any element of the space $\mathcal{M}(\Omega)$. Furthermore, these elements can be made to be absolutely continuous with respect to Lebesgue measure [21]. Specifically, the parameter choice $\alpha_j = cj^2$, with $c > 0$ and $j$ the level of the set in question within the tree, satisfies this condition and ensures that samples from the PT are almost surely continuous. We use this choice throughout our simulations and provide a discussion and robustness analysis for the setting of the constant $c$ in Section 4.3.

The Pólya tree just defined is supported on a one-dimensional domain, but a multi-dimensional extension—in which sets $C \subseteq \Omega^d$ are binary-divided in each of $d$ dimensions simultaneously at each step—is considered by Paddock [27]. In this construction, the children of $C$ are assigned probability mass by means of Dirichlet-distributed random variables $\theta$ supported on the $2^d$-dimensional simplex,

[2]We abuse notation slightly by writing $q$ both for the measure and for its density, the existence of which is always assumed.

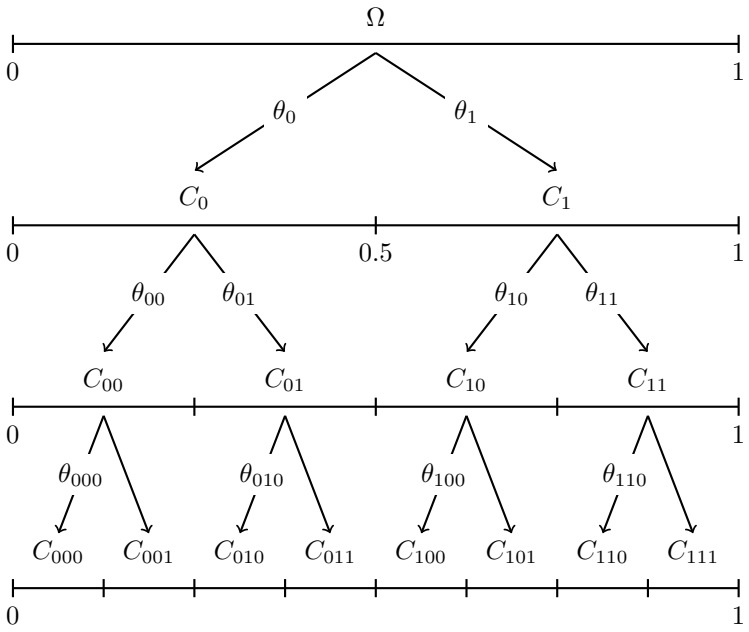


FIG. 1. Construction of a Pólya tree distribution on $\Omega = [0, 1]$. From each set $C_*$, a particle of probability mass passes to the left with (random) probability $\theta_{*0}$ and to the right with probability $\theta_{*1} = 1 - \theta_{*0}$, with all $\theta_*$ being independently Beta-distributed as described in the main text.

generalising the Beta-distributed $\theta$ of the one-dimensional PT.[3] Indices $\varepsilon_i$ now take values in the expanded set $\{0, 1, \ldots, 2^d - 1\}$, we have $(\theta_{\varepsilon^{j-1}0}, \ldots, \theta_{\varepsilon^{j-1}(2^d-1)}) \sim \mathrm{Dir}(\alpha_j, \ldots, \alpha_j)$, and the sets $E^j$ and $E^*$ are redefined accordingly. Note that, by definition, $\sum_{k=0}^{2^d-1} \theta_{\varepsilon^{j-1}k} = 1$.

2.1. **Bayesian inference with Pólya trees.** Pólya trees benefit from the conjugacy of the Binomial and Beta (in the multi-dimensional setting: the Multinomial and Dirichlet) distributions, allowing a simple expression to be derived for the posterior measure over $\mathcal{M}(\Omega)$ after data $X_{1:N} \equiv \{X_1, \ldots, X_N\}$ have been observed. Let $\{\theta\}$ be the collection of all $\theta_{\varepsilon^*}$, and $\Pi_{\Omega^d}$ the set of all $C_{\varepsilon^*}$ arising in the recursive partitioning procedure. Then the density of a point $\mathbf{x} \in \Omega^d$ is given by

$$q(\mathbf{x}|\{\theta\}, \Pi_{\Omega^d}) = \prod_{j=1}^{\infty} \prod_{\varepsilon^j \in E^j} \theta_{\varepsilon^j}^{\mathbb{1}[\mathbf{x} \in C_{\varepsilon^j}]} \tag{8}$$

This equation can be viewed loosely as the limiting case of (7) and unpacked by noting that the conjunction of the product over level-$j$ indices $\varepsilon^j$ and the indicator function in the exponent zeroes all contributions from parameters $\theta_{\varepsilon^j}$ not on the path within the tree that leads to $\mathbf{x}$.

A critical point is that in the classical PT model, exact calculation of quantities such as (8) theoretically requires infinite computation, since the tree is of unlimited depth. It is therefore common in practice to truncate the calculation at a finite tree

[3]Recall that if $\theta_0 \sim \mathrm{Beta}(\alpha, \alpha)$ and $\theta_1 = 1 - \theta_0$, then $\theta \equiv (\theta_0, \theta_1) \sim \mathrm{Dirichlet}(\alpha, \alpha)$.

depth $J$. These truncated (also called 'finite' or 'partially-specified') Pólya trees (TPT) are discussed by Lavine [21] and Mauldin et al. [26]. While full Kullback–Leibler support over $\mathcal{M}(\Omega)$ is no longer guaranteed, bounds on the pointwise error of the posterior measure [21] and $L_1$ error of the predictive density [13] are available. Hanson & Johnson [14] formalise the definition of the TPT by specifying a base measure $\mu$ that the 'leaf' sets at the bottom level $J < \infty$ are taken to follow. If this base measure is uniform then the TPT outputs piecewise-constant measures (ie. random histograms). We follow this approach in the simulations in Section 4, but other base measures can be used—for example if $\Omega$ is unbounded and $\mu$ is taken to be a $d$-dimensional Gaussian measure [13]. The density function for the multivariate TPT is given by

$$q(\mathbf{x}|\{\theta\},\Pi_{\Omega^d},\mu) = \sum_{\varepsilon^J\in E^J} \frac{\mu(\mathbf{x})\mathbb{1}[\mathbf{x}\in C_{\varepsilon^J}]}{\mu(C_{\varepsilon^J})} \prod_{j=1}^{J-1}\prod_{\varepsilon^j\in E^j} \theta_{\varepsilon^j}^{\mathbb{1}[\mathbf{x}\in C_{\varepsilon^j}]} \tag{9}$$

The first fraction in this equation is the normalised base density of the point $\mathbf{x}$ within its level-$J$ set.

We now combine the prior with the likelihood. Conjugacy not only means that the posterior is itself a Pólya tree, but also that the branching variables $\{\theta\}$ can easily be marginalised. Assuming henceforth that $\mu$ is indeed uniform, this gives the TPT marginal likelihood

$$\begin{aligned} p_X(\mathbf{X}_{1:N}|\{\alpha\},\Pi_{\Omega^d},\mu) &= \int \prod_{i=1}^N q(\mathbf{X}_i|\{\theta\},\Pi_{\Omega^d},\mu)p(\{\theta\}|\{\alpha\})\; d\{\theta\} \\ &= \frac{1}{2^{dJn}} \prod_{j=1}^{J-1} \frac{\Gamma(2^d\alpha_j)\cdot\prod_{\varepsilon^j\in E^j}\Gamma(\alpha_j+n_{\varepsilon_j}(\mathbf{X}_{1:N}))}{\Gamma(\alpha_j)^{2^d}\cdot\Gamma(2^d\alpha_j+\sum_{\varepsilon^j\in E^j} n_{\varepsilon_j}(\mathbf{X}_{1:N}))} \end{aligned} \tag{10}$$

Here, $n_{\varepsilon_j}(\mathbf{X}_{1:N})$ counts the number of data $\mathbf{X}_{1:N}$ in the set $\varepsilon_j$. It is then possible to derive the predictive distribution, and using this, an alternative expression for the marginal likelihood that is easier to work with in practice.

$$p_X(\mathbf{x}|\mathbf{X}_{1:N},\{\alpha\},\Pi_{\Omega^d},\mu) = \prod_{j=1}^J \frac{2^d\alpha_j + 2^d n_j(\mathbf{x};\mathbf{X}_{1:N})}{2^d\alpha_j + n_{j-1}(\mathbf{x};\mathbf{X}_{1:N}))} \tag{11}$$

$$p_X(\mathbf{X}_{1:N}|\{\alpha\},\Pi_{\Omega^d},\mu) = \prod_{i=2}^N\prod_{j=1}^J \frac{2^d\alpha_j + 2^d n_j(\mathbf{X}_i;\mathbf{X}_{1:i-1})}{2^d\alpha_j + n_{j-1}(\mathbf{X}_i;\mathbf{X}_{1:i-1})} \tag{12}$$

In these equations, $n_j(\mathbf{x};\mathbf{X}_{1:N})$ counts the number of data in $\{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ that are at the same level-$j$ set as $\mathbf{x}$, ie. $n_j(\mathbf{X}_i;\mathbf{X}_{1:i-1}) = \sum_{k=1}^{i-1}\mathbb{1}[\mathbf{X}_k\in C_{\varepsilon^j}]\mathbb{1}[\mathbf{X}_i\in C_{\varepsilon^j}]$.

2.2. **Pólya tree models for conditional distributions.** In this section we describe how the canonical Pólya tree construction described in Section 2 can be extended to model *conditional* distributions. Doing so first requires a notion of randomised partitioning called 'optional stopping'. This was first proposed by Wong & Ma [39] as an alternative solution to the problem of ensuring that computation time in PT modelling be made almost surely finite. In this paradigm, called the Optional Pólya tree (OPT), the partitioning of $\Omega$ is augmented at each step by the drawing of independent Bernoulli-distributed stopping variables $S$. For a set $C_*$ arising in the partitioning of $\Omega$, if the corresponding $S_*$ is equal to 1 then $C_*$ is divided no further and a uniform distribution is placed on it. If $S_* = 0$ then

a binary split takes place as usual. This outcome of this procedure is a (random) partition of varying granularity across the domain.

As long as the Bernoulli parameter $\rho$ controlling the probability $\Pr(S_* = 1)$ is uniformly greater than 0 for all sets $C_*$, it is easy to see that this algorithm will result in all of $\Omega$ (but for a set of measure zero) being 'stopped' in finite time with probability one. The additional randomness introduced by this partitioning procedure is itself marginalised to give quantities analogous to (11) and (12) that can be calculated in finite time. Given certain further technical conditions, a full-support result akin to that for the classical PT is also available.

The optional stopping principle is then further leveraged in Ma [22, 23], in which multi-scale mixtures of OPTs are used as models for conditional probability distributions; this is called the conditional Optional Pólya tree (cond-OPT) [23]. The basic idea is to construct a random conditional density $q(x|z) \in \mathcal{M}(\mathbb{R} \times \mathbb{R})$ by partitioning the predictor space $\Omega_Z$ using the optional-stopping algorithm described above, then for each set $A$ arising from this procedure to construct an independent (unconditional) OPT random density on the response space $\Omega_X$ but using *only* those data $X_i$ whose corresponding $Z_i$ value lies in $A$. Finally, the multiple independent models over $\Omega_X$ are combined in a weighted sum (with the weights determined by the partition of $\Omega_Z$), giving a random conditional density $q(x|z)$.

The measure constructed this way has full (total variation) support on the space $\mathcal{M}(\mathbb{R} \times \mathbb{R})$ of conditional density functions supported on $\Omega_X \times \Omega_Z$ [23], and as such is a direct generalisation of the unconditional Pólya tree family of models so far discussed, immediately inheriting many of their strengths. This construction for modelling random conditional density functions forms a central part of our work and we describe it in much greater detail in the next section.

## 3. A bayesian conditional independence test.

Recall that we seek to compare the hypotheses $H_0 : X \perp\!\!\!\perp Y \mid Z$ versus $H_1 : X \not\perp\!\!\!\perp Y \mid Z$.

Call the support of $X$, $Y$ and $Z$ respectively $\Omega_X$, $\Omega_Y$ and $\Omega_Z$ and assume that $\Omega := \Omega_X \times \Omega_Y \times \Omega_Z$ is a compact subset of $\mathbb{R}^3$. We will define three nonparametric priors, one for each of the three conditional density functions appearing in (5). Then, by incorporating the data $W$ and marginalising the randomness in the posterior, we will derive the three conditional marginal likelihoods required to calculate the Bayes Factor.

We use $p_{X|Z}$ as our running example; the models for $p_{Y|Z}$ and $p_{XY|Z}$ are the same, with the obvious modifications. The approach consists in first constructing a random partition of $\Omega_Z$ and then, for each partition block $A$, generating the distribution of $X$ conditionally on $Z \in A$ using a truncated Pólya tree (TPT). The first step is to partition $\Omega_Z$ using the optional-stopping binary recursive partitioning procedure described in Section 2.2. This produces a *random* partition of $\Omega_Z$—this is an intrinsic feature of this scheme. This additional randomness will itself be marginalised in order to calculate the conditional marginal likelihood $p_{X|Z}(W)$. Following Ma [23], this is done in practice by constructing a *non-random* binary partition $\Pi_{\Omega_Z}$ and performing a recursive calculation on the resulting tree. We now explain this calculation in detail.

For any $A \in \Pi_{\Omega_Z}$, let $W_A = \{(X_i, Y_i, Z_i) \; ; \; i = 1, \ldots, N : Z_i \in A\}$ be the subset of the data $W$ whose $Z$ component is in $A$, and let $N_A = |W_A|$ be the cardinality of this set. We also write $X_A$ for the set of $X$ components of $W_A$, and similarly for $Y_A$ and $Z_A$. For each set $A$, we consider a 'local' conditional distribution of $X$ given

$Z = z$ (which is assumed to be constant across all $z \in A$) and use a TPT prior for this distribution. The 'local' likelihood of the data $X_A$ is therefore given by

$$q^0_X(A) := \prod_{i=1}^{N_A} q((X_A)_i | \{\theta\}, \Pi_{X,A}) \ ,$$

where the contributions from individual data points are given by (9), and $\Pi_{X,A}$ denotes the partition that 'separates' $X_A$. More precisely, the partition $\Pi_{X,A} \subseteq \Pi_{\Omega_X} (\equiv \Pi_{X,\Omega_Z})$ of $\Omega_X$ is defined such that all leaf sets contain either 0 or 1 data point from $X_A$.[4] The full multi-scale conditional likelihood $q_X(A)$ is then determined recursively by drawing stopping variables $S_X(A)$, and calculating $q^0_X(A)$ for all sets $A$ arising in the resulting random partition of $\Omega_Z$. For any set $A_*$ which remains unstopped, we call its two children $A_{*0}$ and $A_{*1}$. Then $q_X(A_*)$ is given by

$$q_X(A_*) := \begin{cases} q^0_X(A_*) & \text{if } S_X(A_*) = 1, \\ q_X(A_{*0}) q_X(A_{*1}) & \text{if } S_X(A_*) = 0. \end{cases} \tag{13}$$

Equivalently, this can be written as an additive mixture.

$$q_X(A_*) = S_X(A_*) q^0_X(A_*) + (1 - S_X(A_*)) q_X(A_{*0}) q_X(A_{*1}) \tag{14}$$

To calculate the conditional *marginal* likelihood, this expression needs to be integrated to marginalise the randomness from both the local likelihoods $\{q^0_X\}$, and the partitioning procedure, determined by $\{S_X\}$. We write the local marginal likelihoods as $\Phi^0_X(A) := p_X(X_A | \{\alpha\}, \Pi_{X,A})$, and from equation (12) we have

$$\Phi^0_X(A) = \prod_{i=2}^{N_A} \prod_{j=1}^{J_X} \frac{2\alpha_j + 2n_j((X_A)_i; (X_A)_{1:i-1})}{2\alpha_j + n_{j-1}((X_A)_i; (X_A)_{1:i-1})} \tag{15}$$

where $J_X$ is the maximum depth of the partition $\Pi_X$. The complete conditional marginal likelihood $\Phi_X(A) := p_{X|Z}(W_A)$ is then obtained by marginalising the partitioning randomness from (14). Letting $\rho(A_*) = \Pr(S_X(A_*) = 1)$, we have

$$\Phi_X(A_*) = \rho(A_*) \Phi^0_X(A_*) + (1 - \rho(A_*)) \Phi(A_{*0}) \Phi(A_{*1}) \tag{16}$$

This recursion is performed in practice by starting from the leaf sets of the most extensive non-random separating partition $\Pi_{\Omega_Z}$ and applying the following algorithm, until the root $\Omega_Z$ is reached.

$$\Phi_X(A_*) := \begin{cases} \Phi^0_X(A_*) & \text{if } A_* \text{ is a leaf set}, \\ \rho(A_*)\, \Phi^0_X(A_*) + (1 - \rho(A_*)) \Phi_X(A_{*0}) \Phi_X(A_{*1}) & \text{if not.} \end{cases} \tag{17}$$

The value of this function at the root $\Omega_Z$ is the conditional marginal likelihood we require, ie.

$$p_{X|Z}(W) = \Phi_X(\Omega_Z) \ . \tag{18}$$

The variables $\rho(A) \in (0, 1)$ function as mixing parameters and we take them to be constant and equal to 0.5 for all sets $A$—we discuss this choice in Section 4.3. Equation (16) makes clear the way in which the conditional marginal likelihood $\Phi_X(\cdot)$ is formed of a multi-scale additive mixture of TPT marginal likelihoods $\Phi^0_X(\cdot)$.

[4] In practice such partition is calculated most efficiently by constructing the most extensive tree $\Pi_{X,\Omega_Z}$ once, then pruning it to find the $\Pi_{X,A}$ for each $A$.

**Bayesian nonparametric test to assess** $H_0 : X \perp\!\!\!\perp Y \mid Z$ **vs.** $H_1 : X \not\perp\!\!\!\perp Y \mid Z$

**inputs:** data $W = \{(X_i, Y_i, Z_i) : i = 1, \ldots, N\}$; parameters $\rho, c$;
finite 'separating' partitions $\Pi_{\Omega_Z}, \Pi_{\Omega_X}, \Pi_{\Omega_Y}$ and $\Pi_{\Omega_{XY}}$

**for all** $A$ in $\Pi_{\Omega_Z}$
*//partition pruning*
$W_A = \{(X_i, Y_i, Z_i) : Z_i \in A\}$ $(W_A \equiv (X_A, Y_A, Z_A))$
construct $\Pi_{X,A}$, $\Pi_{Y,A}$ and $\Pi_{XY,A}$ by pruning $\Pi_{\Omega_X}, \Pi_{\Omega_Y}$ and $\Pi_{\Omega_{XY}}$,
keeping only those blocks containing $\geq 2$ data points from $W_A$

*// calculate TPT marginal likelihoods* (12)
$\Phi^0_X(A) \leftarrow p_X(X_A|\{\alpha\}, \Pi_{X,A})$
$\Phi^0_Y(A) \leftarrow p_Y(Y_A|\{\alpha\}, \Pi_{Y,A})$
$\Phi^0_{XY}(A) \leftarrow p_{XY}((X_A, Y_A)|\{\alpha\}, \Pi_{XY,A})$

*// calculate conditional marginal likelihoods* (17)
**for all** leaf sets $A$ in $\Pi_{\Omega_Z}$
$\Phi_X(A) \leftarrow \Phi^0_X(A)$
$\Phi_Y(A) \leftarrow \Phi^0_X(A)$
$\Phi_{XY}(A) \leftarrow \Phi^0_{XY}(A)$

**for all** non-leaf sets $A$ in $\Pi_{\Omega_Z}$ with children $A_0$ and $A_1$ (17)
*// traversal order from leaf sets towards root*
$\Phi_X(A) \leftarrow \rho\Phi^0_X(A) + (1-\rho)\Phi_X(A_0)\Phi_X(A_1)$
$\Phi_Y(A) \leftarrow \rho\Phi^0_Y(A) + (1-\rho)\Phi_Y(A_0)\Phi_Y(A_1)$
$\Phi_{XY}(A) \leftarrow \rho\Phi^0_{XY}(A) + (1-\rho)\Phi_{XY}(A_0)\Phi_{XY}(A_1)$

**output:** $\mathrm{BF} \leftarrow \Phi_X(\Omega_Z)\Phi_Y(\Omega_Z)(\Phi_{XY}(\Omega_Z))^{-1}$ (19)

FIG. 2. Pseudocode for the proposed Bayesian nonparametric test for conditional independence

The equivalent calculation is undertaken to find $p_{Y|Z}(W) \equiv \Phi_Y(\Omega_Z)$ and—now using the bivariate version of the TPT—$p_{XY|Z}(W) \equiv \Phi_{XY}(\Omega_Z)$. The Bayes Factor (5) is then given by

$$\mathrm{BF}(H_0, H_1) = \frac{\Phi_X(\Omega_Z)\Phi_Y(\Omega_Z)}{\Phi_{XY}(\Omega_Z)} \tag{19}$$

and, the posterior probability of conditional dependence $p(H_1|W)$ can then be obtained using (4). The algorithm described in this section is summarised in the pseudocode in Figure 2.

4. **Experiments.** In this section we describe some example experiments to elucidate the operation and output of the proposed approach. We stress once again that the output of our algorithm is a Bayesian posterior probability value $p(H_1|W)$ which is directly interpretable as a "probability of conditional dependence", in contrast to previous approaches, which derive or approximate a threshold value for a classical test statistic. This fundamental difference makes direct comparison with existing methods challenging.

4.1. **Synthetic data.** Our first set of experiments uses synthetic datasets constructed by the formulae in the first column of Figure 3. The measures from which the data are sampled are designed in such a way that every combination of unconditional independence/dependence and conditional independence/dependence is

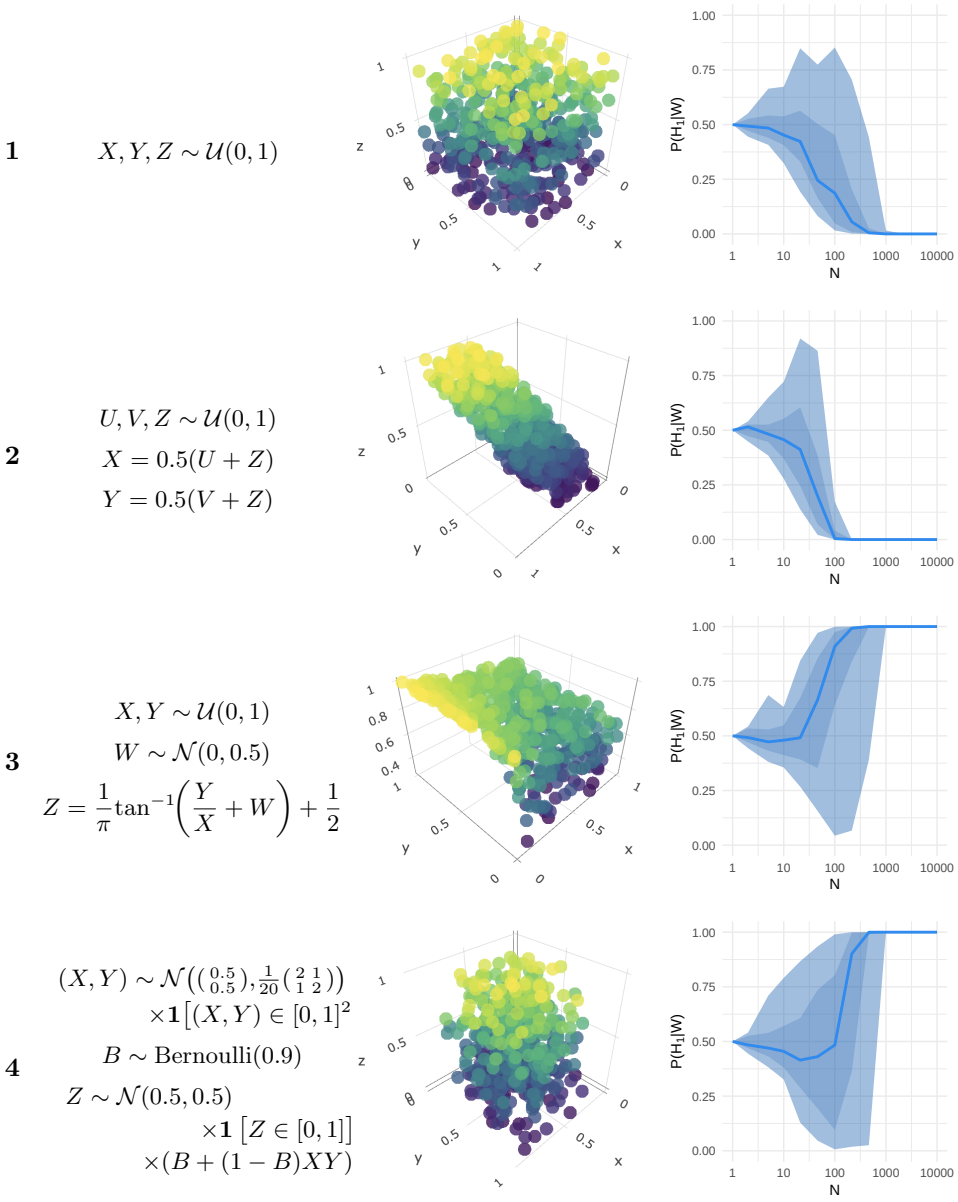


FIG. 3. Application of the proposed Bayesian testing procedure to four synthetic datasets supported on $[0,1]^3$, chosen such that all combinations of unconditional and conditional dependence/independence are represented. The final column gives the ensemble of probabilities of conditional dependence $p(H_1|W)$ output by the test over 100 repetitions at varying values of data size $N$, with the blue line representing the median, and the dark and light shaded regions representing the (25,75)-percentile and (5,95)-percentile ranges respectively.

represented. Specifically, in model 1 it holds that $X \perp\!\!\!\perp Y$ as well as $X \perp\!\!\!\perp Y \,|\, Z$; in model 2 we have $X \not\perp\!\!\!\perp Y$ but $X \perp\!\!\!\perp Y \,|\, Z$; in model 3 it holds that $X \perp\!\!\!\perp Y$ though $X \not\perp\!\!\!\perp Y \,|\, Z$, and in model 4 we have $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \,|\, Z$. In each case, $(X, Y, Z)$ are by construction supported on $\Omega_X = \Omega_Y = \Omega_Z = [0, 1]$. Example 3-dimensional scatter plots are given for each model in the middle column.

We highlight specifically model 4, for which $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \,|\, Z$, though the generating process is a mixture and for 90% of the data it holds that $X \perp\!\!\!\perp Y \,|\, Z$. Noting definition (1) ("for all $z$"), we would like a partial conditional dependence of this type to be detected by a hypothesis test, even if it derives from only a small subset of the data.

We vary the number of data $N$ between 1 and $10^5$, and for each of several values of $N$ in this range we run 100 repetitions of our procedure using datasets generated by different random seeds. We consider binary recursive partitions of $\Omega_X = \Omega_Y = \Omega_Z = [0, 1]$ which at level $j$ have the form

$$[0, 1] = \bigcup_{k=0}^{2^j - 1} \left[ \frac{k}{2^j}, \frac{k+1}{2^j} \right). \tag{20}$$

The maximum tree depths $J_Z$ (in the predictor space) and $J_X$, $J_Y$ and $J_{XY}$ (in the response spaces) are all set at $\lceil \log_2(N) \rceil$, following a widely-used rule of thumb [14]. In addition, we assume an equal prior value for both hypotheses, so that $p(H_0) = p(H_1) = 0.5$.

We plot the range of test outputs $p(H_1|W)$ in the right-hand column of Figure 3, with the blue line representing the median, and the dark- and light-blue shaded regions representing the (25,75)-percentile range and the (5,95)-percentile range respectively. In the low-data limit, the test output $p(H_1|W)$ converges to 0.5 as expected, indicating reversion to the prior probability $p(H_1)$, while for values of $N$ of $10^4$ and greater the test consistently returns a probability value very close to 0 or 1, correctly determining in each case the hypothesis that reflects the ground truth.

In the approximate range $N = 10^1$ to $10^3$, a relatively large uncertainty is present in the output. In the case of the two examples for which $X \not\perp\!\!\!\perp Y \mid Z$ (models 3 and 4), there is a noticeable tendency in this range to falsely favour $H_0$, before $p(H_1|W)$ converges to 1 correctly as $N > 10^4$. This is a manifestation of the natural Occam Factor present in the test, favouring the simpler model $H_0$ where insufficient data exists to conclusively support $H_1$ [24, §28]. The same phenomenon was observed in the unconditional independence testing procedure upon which this work builds [7].

4.2. **Real data.** We now apply our method to a representative real dataset to further illustrate its potential. We consider the California Cooperative Oceanic Fisheries Investigations (CalCOFI) Bottle data, a collection of hydrographic readings from maritime stations off the Californian coast collected over a period of 70 years. These data (available at `calcofi.org`) contain numerous examples of variables with highly non-linear or even non-functional dependence relations. This is illustrated in Figure 4, which shows pairwise scatter plots of representative data from three of the variables in the dataset.

The complete dataset consists of 864,863 observations of 74 variables, but with a high incidence of missing data and numerous strong 'trivial' linear correlations. As a consequence, we first remove all variables for which there is at least one other variable with which it has no common data at all. We then calculate pairwise correlation between the remaining variables, and retain only one representative from

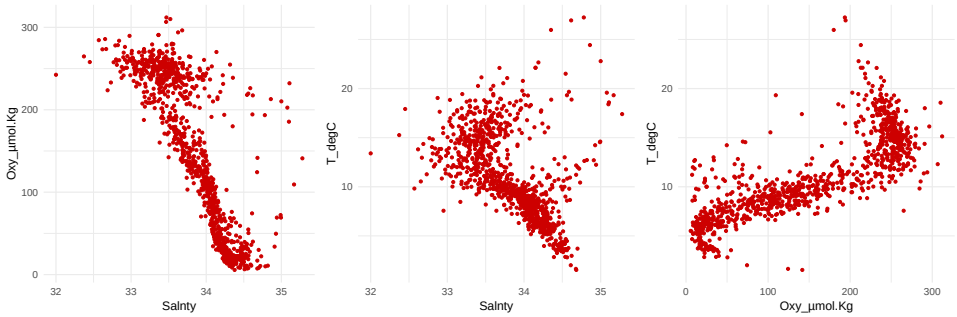


FIG. 4. Marginal scatter plots from the CalCOFI Bottle dataset showing the pairwise relationships between `Salnty`, `Oxy_µmol.Kg` and `T_degC`. The nonlinear nature of the dependences is immediately apparent.

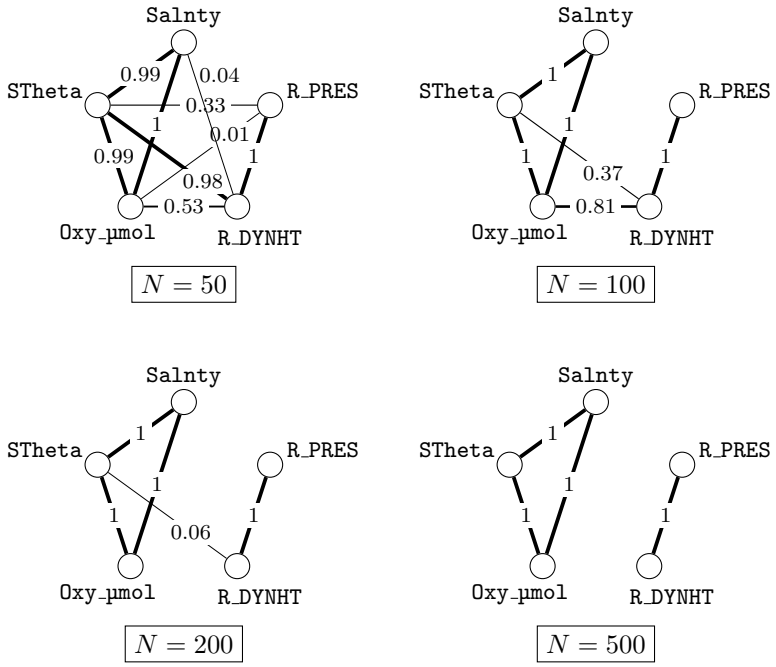


FIG. 5. Example pairwise dependence graphs output by the Bayesian conditional independence test for five variables from the CalCOFI dataset, conditional on `T_degC`, for four different sizes of subsample drawn from the complete dataset. The numbers associated with each edge are the posterior probabilities of conditional dependence $p(H_1|W^{(N)})$ and are given to two decimal places; where no edge is shown, this indicates $p(H_1|W^{(N)}) < 0.005$.

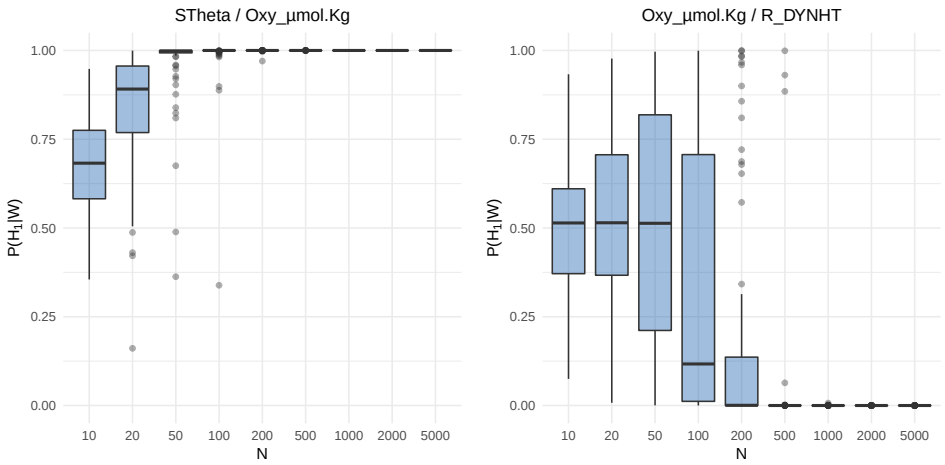


FIG. 6. Box-plots giving the output posterior probability of conditional dependence $p(H_1|W^{(N)})$ for 100 repetitions of the Bayesian conditional independence test applied to randomly-drawn subsamples of various sizes $N$ from the CalCOFI dataset. The left-hand plot gives a representative example of a pair of variables conditionally dependent given `T_degC`, while the right-hand plot gives a representative conditionally independent pair.

groups with pair correlations all greater than 0.99. This leaves 657,216 observations of six variables, these being `T_degC` (Temperature), `Salnty` (Salinity), `STheta` (Potential density), `Oxy_µmol.Kg` (Oxygen in micromoles per kg), `R_DYNHT` (Dynamic height) and `R_PRES` (Pressure).

For the purposes of exposition, we focus on the case where $Z$ is the variable `T_degC`, and $X$ and $Y$ are chosen from the remaining five variables. Though the number of observations remaining even after pre-processing is not significantly lower than in the full dataset, we subsample sets of much smaller cardinality to demonstrate the ability of our test to correctly identify conditional dependence relations in limited data settings. This also serves to effectively eliminate correlation between observations—which would otherwise be strong in this type of time series data—meaning we are able to avoid violating the assumption of i.i.d. data.

For each of the 10 possible pairs $(X, Y)$ chosen from the five remaining variables, we subsample $N = 50$, 100, 200 then 500 observations and denote the resulting partial datasets $W^{(N)}$. Figure 5 gives the graph corresponding to the pairwise dependences found among these five variables conditional on `T_degC` for one example draw of each size of subsample, where the number associated with each graph edge is the posterior probability of dependence $p(H_1|W^{(N)})$. The uncertainty in the existence or otherwise of a conditional dependence relation is reflected in the smaller $N$ cases by the posterior probabilities shown for those edges, which are away from 0 and 1. This type of output would be unavailable with a classical test. By $N = 200$ (and certainly by $N = 500$), the recovered graph emerges clearly. All probabilities are given to two decimal places; where no edge is shown, this indicates $p(H_1|W^{(N)}) < 0.005$.

The graphs in Figure 5 are given as an example to show the nature of the output possible with the use of a Bayesian algorithm for this problem. In Figure 6 we

give aggregated box-plots for 100 repetitions of the above procedure—analogous to the plots in Figure 3—for two example variable pairs. These show, as expected, a range in the output posterior probabilities of dependence for the smaller values of $N$. If some sort of thresholding were implemented to produce the equivalent decision output of a classical test (for example: "reject $H_0$ if $p(H_1|W^{(N)}) > 0.5$, do not reject $H_0$ otherwise"), then this set of test runs could be used as the basis of an empirical power analysis.

This type of output is in many ways more informative than the output of a classical test. As can be seen in Figure 6, the algorithm occasionally returns what appears to be a fully incorrect answer (ie. a posterior probability of 1 when the majority of other runs strongly imply a state of conditional independence). This is the equivalent of a classical Type I error. More often, however, the algorithm returns a probability value strictly between 0 and 1—this output is richer and can be interpreted by the analyst more readily as representing an uncertain test outcome.

4.3. **Implementation.** Practical implementation of the proposed algorithm given in Figure 2 requires the setting of the two hyperparameters $c$ and $\rho$ as well as recursively-constructed partitions for the various sample spaces. The locations of the splits in such partitions is known to affect inference in the Pólya tree family of models [27]. As a default, we suggest two practical approaches to the reader. In the case of a sample space $\Omega$ with compact support, a simple binary partitioning consists of subdividing each set into two subsets of equal size. For $\Omega = [0, 1]$, we thus obtain the partition defined by (20). This is the approach used for the TPT models in the experiments above. Similarly, a quaternary recursive partition of $[0, 1] \times [0, 1]$ can be constructed by subdividing each two-dimensional set into four square quadrants of equal size.

Another approach, which is also suitable for non-compact sample spaces, is to construct a partition based on the quantiles of a pre-defined distribution $G$; a Gaussian distribution is typically used. For our purposes, it is clear that the partitions of $\Omega_X$, $\Omega_Y$ and $\Omega_Z$ should be constructed separately in order to preserve independence relations. The quaternary recursive partition of $\Omega_{XY} = \Omega_X \times \Omega_Y$ is then constructed from the two binary recursive partitions of $\Omega_X$ and $\Omega_Y$. The parameters of the distribution $G$—such as the mean and variance in the case that $G$ is Gaussian—can be derived from empirical estimates of the location and spread of the samples.

The mixing parameter $\rho$ controls the probability of stopping during the partitioning of $\Omega_Z$ and thereby defines the balance between the contributions of the restricted-data marginal likelihoods from different scales in the multi-scale mixture model. We have chosen to keep $\rho$ independent of the set $A$, however there is no theoretical impediment to letting $\rho$ depend on $A$. Wong & Ma [39] and Ma [23] both fix $\rho = 0.5$ for all their simulations and provide no further discussion of it.

The second hyperparameter is the constant $c$ in the level-dependent Dirichlet hyperparameter $\alpha_j = cj^2$. The question of how to set this is present in all work on Pólya trees and is in general open. Berger & Guglielmi [1] write that $c$ "is very difficult to specify", and it is clear that its value can affect inference. Hanson & Johnson [14] point out that in the case of the TPT, the limit $c \to 0$ essentially turns the model into the empirical distribution of the data, while the opposite limit $c \to \infty$ approaches the parametric model defined by the base measure. In practice, $c = 1$ is a common (though ultimately arbitrary) default choice. Other strategies, such as empirical estimation of $c$, have recently been considered [42].

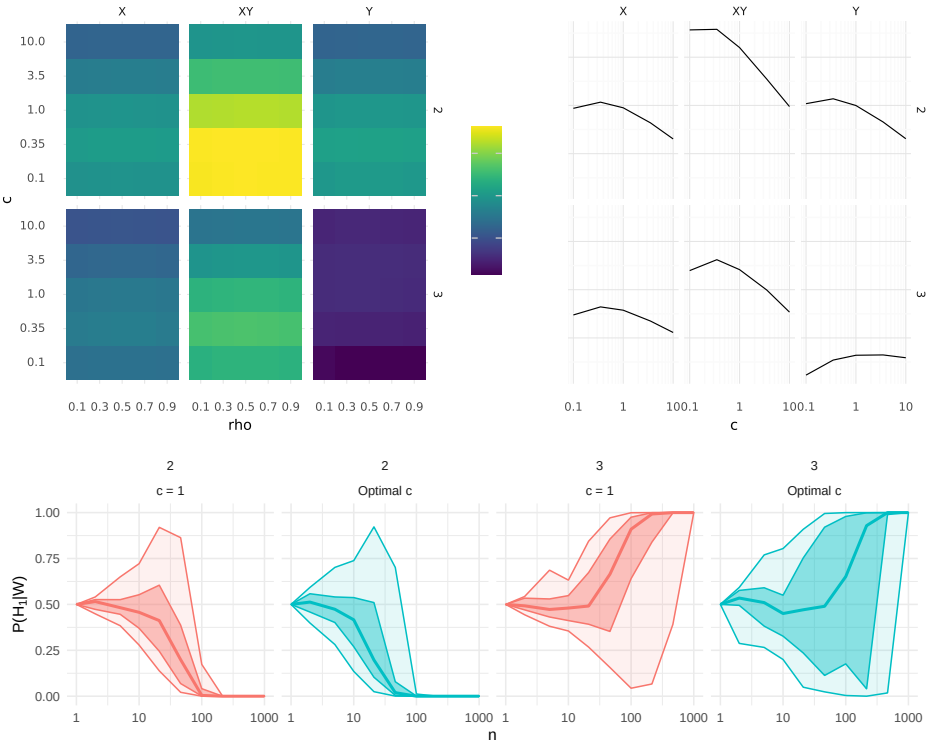


FIG. 7. *Top left:* Heat map of conditional marginal likelihood values for the three constituent models over $\Omega_X$, $\Omega_Y$ and $\Omega_X Y$ for the second and third models of Figure 3. *Top right:* 'Slices' from this heatmap with $\rho = 0.5$. *Bottom:* Test outputs for 100 repetitions of the second and third models of Figure 3. Red plots fix $c = 1$ (output identical to Figure 3), while the blue plots use the optimising values $\hat{c}$ from the plot above.

We ran preliminary studies to gauge the effect that $\rho$ and $c$ have on the output of our test for the second and third example models in Figure 3. Plots of marginal likelihood values, with each of the three models over $\Omega_X$, $\Omega_Y$ and $\Omega_{XY}$ considered separately, are given in Figure 7. From these it is possible to note the relative lack of sensitivity of the conditional marginal likelihood to variations in $\rho$ for values between approximately 0.3 and 0.7. Similar conclusions could be drawn from the remaining two models, not shown here. This evidence, paired with the stated approach of Wong & Ma [39] and Ma [23], justifies our setting $\rho = 0.5$ throughout.

As expected, there is a greater degree of sensitivity amongst the individual marginal likelihood values to the value of $c$. In Figure 7 we contrast the effect on the output posterior probability of conditional independence of setting $c = 1$ throughout (as in Figure 3), and of setting $c$ to the value that maximises the conditional marginal likelihood over a grid of test values for each of the three constituent models separately, ie. $\hat{c}_X \approx \text{argmax}_{c>0}\, p_{X|Z}(W; c)$, and similarly for $\hat{c}_Y$ and $\hat{c}_{XY}$.

The heat map (top left pane of Figure 7) gives the conditional marginal likelihood values for the three constituent models over $\Omega_X$, $\Omega_Y$ and $\Omega_X Y$ for the second and third models of Figure 3. The line plots (top right pane) are 'slices' from this heatmap which fixes $\rho = 0.5$ and seeks to identify the optimal value of $c$. We have

left the vertical scale off these plots since we are only interested in maxima rather than the actual values of the conditional marginal likelihood.

The four panes at the bottom of Figure 7 contrast the test output resulting from the two different approaches to setting $c$, using the same quantile bands as in Figure 7. The red plots (identical to those appearing in Figure 7) fix $c = 1$, while the blue plots use the optimising values from the plot above. Our empirical findings are that, while the value of $c$ does impact the algorithm output, the consistency of the test procedure does not appear to be affected in the large data limit. Theoretical investigation of this assertion would be a fruitful subject for future research.

A more detailed study of the robustness of derived quantities of Pólya trees to changes in hyperparameters is beyond the scope of the present work. For practitioners we recommend either a 'rule of thumb' approach similar to that we have implemented, possibly with a small number of test runs to calibrate, or a more detailed (but correspondingly more time-consuming) set of pre-simulations. The choice will necessarily be dependent on the dataset under consideration and the balance between speed and accuracy called for by the particular use case.

## 5. **Conclusions & discussion.**

In this article we have defined and demonstrated a new Bayesian nonparametric approach to quantifying the relative evidence in favour of independence or dependence of two random variables conditionally on a third. We have done so in a manner that minimises the assumptions required on the unknown joint distribution of $(X, Y, Z)$, by modelling various of its conditional distributions using Pólya trees.

We believe this approach has the potential to be developed in numerous directions, and we hope it will in this way increasingly find application in practical analyses. In its current form, the procedure we describe comes with relatively high computational cost, due primarily to the recursive calculations required, though we hope the line of research opened up by these ideas will soon lead to more efficient implementations. An extension to the multi-dimensional setting, particularly for the conditioning variable $Z$, would be of real use and is the subject of current work.

The Bayesian approach our procedure takes provides a framework in which *both* the hypotheses of conditional independence and conditional dependence can be positively evidenced from a given dataset, unlike the inherently asymmetric hypothesis tests of classical statistics. This is of great importance for causal discovery.

The output of the procedure is a value in the range $[0, 1]$ which can directly be interpreted as a posterior probability of conditional dependence $p(H_1|W)$. This is in notable contrast to previous approaches, even those that work partly within the Bayesian paradigm. This type of output attaches a notion of uncertainty to the result of the test, something absent in classical hypothesis testing. This uncertainty may be propagated further down the 'pipeline' of computation if the test is used as a constituent part of a larger procedure.

Our method also allows substantive prior information on the plausibility of an association to be trivially incorporated, something particularly useful when screening large biological datasets. Lastly, the ability to detect dependences of a highly nonlinear or even non-functional nature allows for much greater confidence in the robustness of any inference procedure into which this type of test is embedded.

## REFERENCES

[1] J. O. Berger and A. Guglielmi, Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives, *J. Amer. Statist. Assoc.*, **96** (2001), 174–184.

[2] W. Bergsma, Testing conditional independence for continuous random variables, *Report Eurandom*, 2004.
[3] T. B. Berrett, Y. Wang, R. F. Barber and R. J. Samworth, The conditional permutation test for independence while controlling for confounders, *J. R. Stat. Soc. B*, **82** (2020), 175–197.
[4] E. Candès, Y. Fan, L. Janson and J. Lv, Panning for gold: Model-X knockoffs for high dimensional controlled variable selection, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **80** (2018), 551–577.
[5] G. Doran, K. Muandet, K. Zhang and B. Schölkopf, A permutation-based kernel conditional independence test, *Proc. 30th Conf. UAI*, 132–141.
[6] M. Escobar and M. West, Bayesian density estimation and inference using mixtures, *J. Amer. Statist. Assoc.*, **90** (1995), 577–588,
[7] S. Filippi and C. Holmes, A Bayesian nonparametric approach to testing for dependence between random variables, *Bayesian Anal.*, **12** (2017), 919–938.
[8] R. Fisher, The distribution of the partial correlation coefficient, *Metron*, **3** (1924), 329–332.
[9] K. Fukumizu, A. Gretton, X. Sun and B. Schölkopf, Kernel measures of conditional dependence, *Adv. Neural Inf. Process. Syst.*, **20**, 489–496.
[10] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, 44. Cambridge University Press, Cambridge, 2017.
[11] J. K. Ghosh and R. V. Ramamoorthi, *Bayesian Nonparametrics*, Springer-Verlag, New York, 2003.
[12] P. Giudici, Bayes factors for zero partial covariances, *J. Statist. Plann. Inference*, **46** (1995), 161–174.
[13] T. E. Hanson, Inference for mixtures of finite Pólya tree models, *J. Amer. Statist. Assoc.*, **101** (2006), 1548–1565.
[14] T. Hanson and W. O. Johnson, Modeling regression error with a mixture of Pólya trees, *J. Amer. Statist. Assoc.*, **97** (2002), 1020–1033.
[15] N. Harris and M. Drton, PC algorithm for nonparanormal graphical models, *J. Mach. Learn. Res.*, **14** (2013), 3365–3383.
[16] P. Hoyer, D. Janzing, J. Mooij, J. Peters and B. Schölkopf, Nonlinear causal discovery with additive noise models, *Adv. Neural Inf. Process. Syst.* **21**, 689–696.
[17] T.-M. Huang, Testing conditional independence using maximal nonlinear conditional correlation, *Ann. Statist.*, **38** (2010), 2047–2091.
[18] R. E. Kass and A. E. Raftery, Bayes factors, *J. Amer. Statist. Assoc.*, **90** (1995), 773–795.
[19] T. Kunihama and D. B. Dunson, Nonparametric Bayes inference on conditional independence, *Biometrika*, **103** (2016), 35–47.
[20] M. Lavine, Some aspects of Pólya tree distributions for statistical modelling, *Ann. Statist.*, **20** (1992), 1222–1235.
[21] M. Lavine, More aspects of Pólya tree distributions for statistical modelling, *Ann. Statist.*, **22** (1994), 1161–1176.
[22] L. Ma, Adaptive testing of conditional association through recursive mixture modeling, *J. Amer. Statist. Assoc.*, **108** (2013), 1493–1505.
[23] L. Ma, Recursive partitioning and multi-scale modeling on conditional densities, *Electron. J. Stat.*, **11** (2017), 1297–1325.
[24] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.
[25] D. Margaritis, Distribution-free learning of bayesian network structure in continuous domains, *Proc. 20th Nat. Conf. Artificial Intel.*, (2005), 825–830.
[26] R. D. Mauldin, W. D. Sudderth and S. C. Williams, Pólya trees and random distributions, *Ann. Statist.*, **20** (1992), 1203–1221.
[27] S. M. Paddock, *Randomized Pólya Trees: Bayesian Nonparametrics for Multivariate Data Analysis*, Thesis (Ph.D.)–Duke University. 1999.
[28] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.
[29] J. Peters, D. Janzing and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, Cambridge, MA, 2017.
[30] J. Peters, J. Mooij, D. Janzing and B. Schölkopf, Causal discovery with continuous additive noise models, *J. Mach. Learn. Res.*, **15** (2014), 2009–2053.
[31] J. Ramsey, A scalable conditional independence test for nonlinear, non-Gaussian data, arXiv:1401.5031.

[32] J. Runge, Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, arXiv:1709.01447.
[33] F. Saad and V. Mansinghka, Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric Bayes, *Proc. Mach. Learn. Res.*, **46** (2017), 632–641.
[34] R. Shah and J. Peters, The hardness of conditional independence testing and the generalised covariance measure, arXiv:1804.07203.
[35] P. Spirtes and C. Glymour, An algorithm for fast recovery of sparse causal graphs, *Soc. Sci. Comput. Rev.*, **9** (1991), 62–72.
[36] E. Strobl, K. Zhang and S. Visweswaran, Approximate kernel-based conditional independence tests for fast non-parametric causal discovery, *J. Causal Inference*, (2019), 20180017.
[37] L. Su and H. White, A consistent characteristic function-based test for conditional independence, *J. Econom.*, **141** (2007), 807–834.
[38] L. Su and H. White, A nonparametric Hellinger metric test for conditional independence, *Econom. Theory*, **24** (2008), 829–864,
[39] W. H. Wong and L. Ma, Optional Pólya tree and Bayesian inference, *Ann. Statist.*, **38** (2010), 1433–1459.
[40] Q. Zhang, S. Filippi, S. Flaxman and D. Sejdinovic, Feature-to-feature regression for a two-step conditional independence test, *Proc. 33rd Conf. UAI*, 2017.
[41] K. Zhang, J. Peters, D. Janzing and B. Schölkopf, Kernel-based conditional independence test and application in causal discovery, arXiv:1202.3775.
[42] J. Zhang, L. Yang and X. Wu, Pólya tree priors and their estimation with multi-group data, *Stat. Pap.*, **60** (2019), 499–525.

*E-mail address*: o@teymur.uk
*E-mail address*: s.filippi@imperial.ac.uk