Electronic Journal of Statistics
Vol. 10 (2016) 3338–3354
ISSN: 1935-7524
DOI: 10.1214/16-EJS1171

Scalable Bayesian nonparametric measures for exploring pairwise dependence via Dirichlet Process Mixtures

Sarah Filippi, Chris C. Holmes

Department of Statistics
University of Oxford
England
e-mail: sarah.filippi@stats.ox.ac.uk; cholmes@stats.ox.ac.uk

and

Luis E. Nieto-Barajas

Department of Statistics
ITAM
Mexico
e-mail: lnieto@itam.mx

Abstract: In this article we propose novel Bayesian nonparametric methods using Dirichlet Process Mixture (DPM) models for detecting pairwise dependence between random variables while accounting for uncertainty in the form of the underlying distributions. A key criteria is that the procedures should scale to large data sets. In this regard we find that the formal calculation of the Bayes factor for a dependent-vs.-independent DPM joint probability measure is not feasible computationally. To address this we present Bayesian diagnostic measures for characterising evidence against a "null model" of pairwise independence. In simulation studies, as well as for a real data analysis, we show that our approach provides a useful tool for the exploratory nonparametric Bayesian analysis of large multivariate data sets.

Keywords and phrases: Bayes nonparametrics, contingency table, dependence measure, hypothesis testing, mixture model, mutual information.

Received December 2015.

1. Introduction

Identifying dependences among pairs of random variables measured on the same sample, producing datasets of the form $\mathrm{D} = \{(x_i, y_i),\ i = 1, \ldots, n\}$, is an important task in modern exploratory data analysis where historically the Pearson correlation coefficient and the Spearman's rank correlation have been used. More recently there has been a move to the use of non-linear or distribution free methods such as those based on Mutual Information (MI) (Cover and Thomas,

2012; Kinney and Atwal, 2014). In this paper we present Bayesian nonparametric methods for screening large data sets for possible pairwise associations (dependencies). Having an explicit probability measure of dependences has numerous advantages both in terms of interpretability and for integration across different experimental conditions and/or within a formal decision theoretic analysis. As data sets become ever larger and more complex we increasingly require Bayesian procedures that can scale to modern applications and this will be a key design criteria here. The main building block of our procedures will be the Dirichlet Process Mixture (DPM) model, which is the most popular Bayesian nonparametric model.

We frame the problem of screening for evidence of pairwise dependence as a nonparametric model choice problem with alternatives:

$$\begin{aligned} \mathcal{M}_0 &: \text{X and Y are independent random variables} \\ \mathcal{M}_1 &: \text{X and Y are dependent random variables .} \end{aligned} \tag{1}$$

Given a set of measurement pairs D, for n exchangeable observations one could then evaluate the posterior probability for competing models $\mathrm{P}(\mathcal{M}_1|\mathrm{D}) = 1 - \mathrm{P}(\mathcal{M}_0|\mathrm{D})$ or consider the Bayes factor $\mathrm{P}(\mathrm{D} \mid \mathcal{M}_0)/\mathrm{P}(\mathrm{D} \mid \mathcal{M}_1)$ which is a measure of the strength of evidence for independence between the two samples against dependence. However with p measurement variables under study there are $\approx \frac{1}{2}p^2$ such pairwise Bayes factors to compute, where even just one such evaluation might be problematic to compute. This motivates us to explore scalable alternatives to a formal Bayesian testing approach, by deriving summary statistics and functionals of the posterior that can provide strong indication in favour or against independence.

Bayesian nonparametric hypotheses testing via Polya tree priors has been the focus of a couple of recent research papers (Holmes et al., 2015; Filippi and Holmes, 2015). Here, however, we specify model uncertainty in the distribution of X and Y via DPMs of Gaussians. This provides flexibility while also encompassing smoothness assumptions on the underlying joint distributions. Another advantage is that DPMs have been widely studied in the Bayesian nonparametric literature with excellent open source implementation packages available (e.g. Jara et al., 2011). Moreover, although not explored here, the use of DPMs makes our approach readily extendable to situations when X and Y are themselves collections of multivariate measurements. Here we consider pairwise dependence between univariate measurements where for $\mathcal{M}_0$, the independence model, the joint distribution factorises into a product of two univariate DPMs on X and Y, while for $\mathcal{M}_1$ we can define a joint DPM model on the bivariate measurement space (X, Y).

In theory, given a DPM prior on the unknown densities, the Bayes factor can be calculated via the marginal likelihood. However this requires integrating over an infinite dimensional parameter space that does not have a tractable form. Moreover, using computational approaches to approximate the marginal likelihood is highly non-trivial, particularly when considering the need to scale to many thousands of comparisons with large p. To overcome this issue we present

two new approaches to deriving scalable diagnostic measures corresponding to probabilistic measures of dependence, bypassing the need to calculate Bayes Factors that might not be feasible or desirable. Our methods are motivated by two recent proposals in the literature (Lock and Dunson, 2013; Kamary et al., 2014), although neither of these papers consider the problem we address here as outlined below.

Our first approach utilises the well known latent allocation, or clustering, structure of the DPM model to induce a partition of the two-dimensional data space. By running a Gibbs sampler under the independence model the cluster allocation of observations to specific mixture components at each iteration can then be used to define a latent contingency table given by the mixture component memberships. For each of these contingency tables we perform a parametric Bayesian independence-vs.-dependence test using conjugate multinomial-Dirichlet priors that lead to explicit analytic forms for the conditional marginal likelihoods. This proposal follows a similar idea considered in Lock and Dunson (2013) who studied the two-sample testing problem. A key difference in what we present here, in addition to that we consider the problem of pairwise dependence, is that Lock and Dunson (2013) use a finite mixture model to induce a partition instead of an infinite nonparametric mixture model used here.

In our second approach, we adapt a recent procedure of (Kamary et al., 2014), turning the model choice problem into an estimation problem by writing the competing models under a hierarchy that incorporates both models, $\mathcal{M}^* = \pi\mathcal{M}_1 + (1-\pi)\mathcal{M}_0$. We investigate the specification of $\mathcal{M}^*$ either as a mixture model with mixing component $0 \leq \pi \leq 1$, or as a predictive linear ensemble of the two sub-models with constraints on the weights. We then estimate π which becomes a measure of the evidence for dependence. DPMs are used to obtain the likelihood associated to each of the competing models in $\mathcal{M}^*$, requiring a separate MCMC run for each potential pair of random variables.

We compare and contrast the two procedures with particular regard to their scalability to large data sets. This latter feature naturally includes the amenity of the methods to simulation with modern parallel computation. We demonstrate that our association measures are scalable and successfully detect some highly non-linear dependences with equivalent performance to the current best conventional methods using mutual information, with the added advantages that fully probabilistic Bayesian methods enjoy. As mentioned above, some of these key advantages includes the ability to integrate results within a formal decision analysis framework, or within optimal experimental design, and the combination of results with other sources of information, or across studies such as arise in meta-analysis.

The rest of the paper is as follows. In Section 2 we review the Dirichlet Process and the DPM of Gaussians. In Section 3 we describe the two approaches to quantify the evidence for dependence using Dirichlet Process Mixtures. In Section 4 we illustrate our approach on the exploratory analysis of a real-world example from the World Health Organisation data set of country statistics and also on simulated data generated from simple models. We conclude the paper with a short discussion in Section 5.

2. Dirichlet Process Mixtures

The Dirichlet process (Ferguson, 1973) is the most important process prior in Bayesian nonparametric statistics. It is flexible enough to approximate (in the sense of weak convergence) any probability law, although the paths of the process are almost surely discrete (Blackwell and MacQueen, 1973). Many years ago this discreteness was considered a drawback but nowadays it is simply a feature that characterises the Dirichlet process. This feature has recently been highly exploited in clustering applications (e.g. (Dahl, 2006)).

The Dirichlet process is defined as follows. Let G be a probability measure defined on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{B}$ the corresponding Borel's σ-algebra. Let G be a stochastic process indexed by the elements of $\mathcal{B}$. G is a Dirichlet process with parameters c and G_0 if for every measurable partition $(B_1, \ldots, B_k)$ of $\mathcal{X}$,

$$(G(B_1), \ldots, G(B_k)) \sim \text{Dir}(cG_0(B_1), \ldots, cG_0(B_k)).$$

From here we can see that, for every $B \in \mathcal{B}$, $\text{E}\{G(B)\}=G_0(B)$ and $\text{Var}\{G(B)\}= G_0(B)\{1 - G_0(B)\}/(c + 1)$. Therefore the parameter c is known as precision parameter and G_0 as the centering measure.

The Dirichlet process when used as a prior induces exchangeability in the data. In notation, let $X_1, \ldots, X_n$ be a sample of random variables such that conditional on G, $X_i \mid G \overset{\text{iid}}{\sim} G$. If we further take $G \sim \mathcal{DP}(c, G_0)$ then the marginal distribution of the data $(X_1, \ldots, X_n)$ once the process G has been integrated out, is characterised by what is known as the Pólya urn (Blackwell and MacQueen, 1973). We start with $X_1 \sim G_0$ then

$$X_n \mid X_1, \ldots, X_{n-1} \sim \frac{cG_0 + \sum_{j=1}^{n-1} \delta_{X_j}}{c + n - 1}. \tag{2}$$

Instead of placing the Dirichlet process prior directly on the observable data, it can be used as the law of the parameters of another model (kernel) that generated the data. In notation, let us assume that for each $i = 1, \ldots, n$,

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} f(x_i \mid \theta_i),$$

with f a parametric density function. We can further take

$$\theta_i \mid G \overset{\text{iid}}{\sim} G$$

with

$$G \sim \mathcal{DP}(c, G_0).$$

This hierarchical specification can be seen as a mixture of density kernels $f(x \mid \theta)$ with mixing distribution coming from a Dirichlet process, i.e., $\int f(x \mid \theta) G(\text{d}\theta)$. This model is known as Dirichlet process mixture (DPM) model and was first introduced by Lo et al. (1984) in the context of density estimation and written in hierarchical form by Ferguson (1983).

The most typical choice of kernel f is the (multivariate) normal, in which case $\theta_i = (\mu_i, \sigma_i^2)$, with scalars mean and variance, in the univariate case, and $\theta_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with mean vector and variance-covariance matrix, in the multivariate case. We will work with this specific kernel throughout this paper.

As can be seen by construction, in the mixture case, the Dirichlet process induces a joint distribution on the set $(\theta_1, \ldots, \theta_n)$ that allows for ties in the θ_i's. This in turn induces a clustering structure in the θ_i's (and X_i's). Posterior inference of the DPM model usually relies on a Gibbs sampler (Smith and Roberts, 1993). At each iteration of the Gibbs sampler the model produces a different clustering structure. The number of clusters is a function of the sample size n and the precision parameter c of the underlying Dirichlet process. The larger the value of c, the larger the number of clusters induced. This clustering structure and parameter c will play a central role in one of the independence test procedures that will be described later.

3. Two approaches for measuring dependence

As noted in Section 1, the calculation or approximation of the formal Bayes factor under $\mathcal{M}_0$ and $\mathcal{M}_1$ is not feasible when considering a large number of model comparisons. Indeed it may not even be desirable given that our objective is to highlight potential departures from independence rather than answer a formal model choice question. In this section we describe two distinct approaches for comparing models $\mathcal{M}_0$ and $\mathcal{M}_1$ defined in (1) based on DPM models that are computable and scalable to large data.

3.1. Contingency tables approach

The first approach is motivated by the paper from Lock and Dunson (2013) who turned a two-sample testing problem into a discrete test on the clustered data. Recall that the two-sample testing problem considers the same measurement variable recorded on separate subjects under two different conditions; whereas we are considering different measurement variables recorded on the same subject. Similar to Lock and Dunson (2013), our procedure consists in marginally discretising the data into ordered categories and performing a Dirichlet-multinomial independence test on the induced contingency table. This amounts to first clustering the data under $\mathcal{M}_0$ and then exploring for evidence of departure from $\mathcal{M}_0$, toward $\mathcal{M}_1$, by testing for statistical association between the cluster memberships in X and Y. Uncertainty in the cluster memberships is accounted for by the DPM defined under $\mathcal{M}_0$, as outlined below.

To begin assume that the data are marginally clustered in K_X and K_Y clusters and denote by $\xi_{X,i} \in \{1 \ldots, K_X\}$ and $\xi_{Y,i} \in \{1, \ldots, K_Y\}$ the cluster indicators for the data points x_i and y_i respectively, for $i = 1, \ldots, n$. Using these cluster indicators, we can construct a contingency table $\mathbf{M}_{\xi_X, \xi_Y} = \{m_{kl}\}$ of size $K_X \times K_Y$, such that $m_{kl} = \sum_{i=1}^{n} I(\xi_{X,i} = k, \xi_{Y,i} = l)$, for $k = 1, \ldots K_X$ and $l = 1, \ldots, K_Y$. The contingency table $\mathbf{M}_{\xi_X, \xi_Y}$ represents a discretised version

of the (unnormalised) marginals and joint distribution of the continuous vector (X, Y). We can then apply Bayesian independence tests for discrete / categorical variables following Gunel and Dickey (1974) and Good and Crook (1987) who proposed a conjugate multinomial-Dirichlet independence test which is described as follows. Let $\mathbf{M}_{\xi_X,\xi_Y} \sim \text{Mult}(n, \mathbf{p})$ with $\mathbf{p} = \{p_{kl}\}$ the matrix of cell probabilities of dimension $K_X \times K_Y$. Consider a conjugate prior distribution $\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = \{\alpha_{kl}\}$ such that $\sum_{kl} \alpha_{kl} = a$. In practice we suggest to use $\alpha_{kl} = a(K_X K_Y)^{-1}$ or $\alpha_{kl} = 1/2$ for all $1 \leq k \leq K_X$ and $1 \leq l \leq K_Y$. Under model $\mathcal{M}_1$ the probability of having observed the counts in $\mathbf{M}_{\xi_X,\xi_Y}$ is

$$\begin{aligned} \mathrm{P}(\mathbf{M}_{\xi_X,\xi_Y} \mid \mathcal{M}_1, \xi_X, \xi_Y) &= \int \mathrm{P}(\mathbf{M}_{\xi_X,\xi_Y} \mid \mathbf{p}) f(\mathbf{p}) \, \mathrm{d}\mathbf{p} \\ &= \frac{\Gamma(a)}{\Gamma(a+n)} \prod_{k,l} \frac{\Gamma(\alpha_{kl} + m_{kl})}{\Gamma(\alpha_{kl})}. \end{aligned} \tag{3}$$

Under the independent model $\mathcal{M}_0$ the observed counts $\mathbf{M}_{\xi_X,\xi_Y}$ can be expressed in terms of the marginal counts $\mathbf{m}_X = \{m_{k\cdot}\}$ and $\mathbf{m}_Y = \{m_{\cdot l}\}$ whose implied distributions are again multinomial with probability vectors $\mathbf{p}_X = \{p_{k\cdot}\}$ and $\mathbf{p}_Y = \{p_{\cdot l}\}$, respectively, with $m_{k\cdot} = \sum_l m_{kl}$, $m_{\cdot l} = \sum_k m_{kl}$, $p_{k\cdot} = \sum_l p_{kl}$ and $p_{\cdot l} = \sum_k p_{kl}$. The induced prior distributions are also Dirichlet with parameters $\boldsymbol{\alpha}_X = \{\alpha_{k\cdot}\}$ and $\boldsymbol{\alpha}_Y = \{\alpha_{\cdot l}\}$. Then, the probability of $\mathbf{M}_{\xi_X,\xi_Y}$ under $\mathcal{M}_0$ becomes

$$\begin{aligned} \mathrm{P}(\mathbf{M}_{\xi_X,\xi_Y} \mid \mathcal{M}_0, \xi_X, \xi_Y) &= \int \mathrm{P}(\mathbf{m}_X \mid \mathbf{p}_X) f(\mathbf{p}_X) \, \mathrm{d}\mathbf{p}_X \int \mathrm{P}(\mathbf{m}_Y \mid \mathbf{p}_Y) f(\mathbf{p}_Y) \, \mathrm{d}\mathbf{p}_Y \\ &= \frac{\Gamma^2(a)}{\Gamma^2(a+n)} \prod_k \frac{\Gamma(\alpha_{k\cdot} + m_{k\cdot})}{\Gamma(\alpha_{k\cdot})} \prod_l \frac{\Gamma(\alpha_{\cdot l} + m_{\cdot l})}{\Gamma(\alpha_{\cdot l})}, \end{aligned} \tag{4}$$

where $\alpha_{k\cdot} = \sum_l \alpha_{kl}$ and $\alpha_{\cdot l} = \sum_k \alpha_{kl}$.

To compare evidence in favour of each model, we use expressions (3) and (4) to compute the Bayes factor $BF_\xi = \mathrm{P}(\mathbf{M}_{\xi_X,\xi_Y} \mid \mathcal{M}_0, \xi_X, \xi_Y)/\mathrm{P}(\mathbf{M}_{\xi_X,\xi_Y} \mid \mathcal{M}_1, \xi_X, \xi_Y)$. Using equal prior probabilities for both models, i.e. $\mathrm{P}(\mathcal{M}_0) = \mathrm{P}(\mathcal{M}_1) = 0.5$, we obtain that the posterior probabilities for the independence and dependence models are $\mathrm{P}(\mathcal{M}_1 \mid \mathbf{M}_{\xi_X,\xi_Y}) = 1/(1 + BF_{\xi_X,\xi_Y}) = 1 - \mathrm{P}(\mathcal{M}_0 \mid \mathbf{M}_{\xi_X,\xi_Y})$ where

$$BF_{\xi_X,\xi_Y} = \frac{\Gamma(a)}{\Gamma(a+n)} \prod_k \frac{\Gamma(\alpha_{k\cdot} + m_{k\cdot})}{\Gamma(\alpha_{k\cdot})} \prod_l \frac{\Gamma(\alpha_{\cdot l} + m_{\cdot l})}{\Gamma(\alpha_{\cdot l})} \prod_{k,l} \frac{\Gamma(\alpha_{kl})}{\Gamma(\alpha_{kl} + m_{kl})}. \tag{5}$$

It should also be noted that this contingency table approach would also afford a conditional frequentist test. For example, consider Pearson's chi-squared test of independence (Pearson, 1922). Under the null hypothesis $\mathcal{M}_0$ of independence, the well known test statistic

$$T = \sum_{k=1}^{K_X} \sum_{l=1}^{K_Y} \frac{(m_{kl} - m_{k\cdot} m_{\cdot l}/n)^2}{m_{k\cdot} m_{\cdot l}/n} \tag{6}$$

follows a χ^2 distribution with $(K_X - 1)(K_Y - 1)$ degrees of freedom. If the test statistic is improbably large according to that chi-square distribution, then one rejects the null hypothesis $\mathcal{M}_0$ in favour of the dependence hypothesis $\mathcal{M}_1$.

The hypothesis testing approach described in this section assumes that the data are marginally clustered. However, these clusters are not known a priori. A Bayesian approach for data clustering is to define a prior distribution over the clustering and then update the posterior based on the evidence provided by the data. Here we make use of the DPM model structure to create an empirical partition of the two-dimensional data space, taking into account the uncertainty on the allocation process. More precisely, we consider two independent DPM prior models for each of the marginal densities with the following specifications:

$$f_{0,X}(x) \sim \int \mathrm{N}(x \mid \theta_X) G_X(\mathrm{d}\theta_X) \quad \text{and} \quad f_{0,Y}(y) \sim \int \mathrm{N}(y \mid \theta_Y) G_Y(\mathrm{d}\theta_Y), \tag{7}$$

where $\theta_X = (\mu_X, \sigma_X^2)$ and $\theta_Y = (\mu_Y, \sigma_Y^2)$, with

$$G_X \sim \mathcal{DP}(c_0, G_0) \ \text{ and } \ G_Y \sim \mathcal{DP}(c_0, G_0) \tag{8}$$

and $G_0 = \mathrm{N}(\mu \mid \mu_0, \sigma^2/k_0)\,\mathrm{IGa}(\sigma^2 \mid \nu/2{-}1, \psi/2)$. The latent clustering structure induced by the DPM models defined by (7) and (8) can then be used to construct a contingency table as described above. Note that in an ideal world one would carefully specify subjective beliefs on the prior marginals for X and Y. However, when the number of variables is large this is not feasible and we require some default specification as done here, by assuming a common prior after suitable transformation of the data.

Although it is clear from the properties of the Dirichlet Process that it induces a partition, in practice it is not easy to determine an optimal one. Fitting a DPM model via a Gibbs sampler provides a partition at each iteration. We can proceed in two different ways. One is to use all potential partitions coming from the MCMC, and for each of them perform the Bayesian independence test and report the expected posterior probability. More precisely, the functional we consider is

$$p_{\mathrm{dep}} = \int \frac{1}{(1 + BF_{\xi_X, \xi_Y})} p(\xi_X, \xi_Y) d\xi_X d\xi_Y \ . \tag{9}$$

This is the procedure we recommend and develop below. An alternative approach would be to consider the selection of one of the partitions using an appropriate optimization criterion, for example using the criterion of Dahl (2006) who proposes to choose the partition that minimises the squared deviations with respect to the average pairwise clustering matrix, and use that single partition to perform the test, ignoring the uncertainty in the partition structure as in Lock and Dunson (2013) for the two-sample test. In *Supplementary Material* we provide an empirical comparison between both procedures.

In the rest of the paper we will focus on the first alternative that considers all potential partitions; we will refer to this procedure as CT-BF – see Algotithm 1.

Algorithm 1 Independence measure based on Contingency table (CT-BF)

Require: Data $\mathrm{D} = \{x_i, y_i\}_{i=1}^n$
Require: Prior parameters a
Require: Prior parameters for the DPM and number of iterations N_{it}
Ensure: Probability of dependence p_{dep}

DPM inference:
Infer a DPM model for the distribution $f_{0,X}(x)$ using a Gibbs Sampler with n_{it} iterations
$\rightarrow$ for each iteration $1 \leq j \leq N_{\mathrm{it}}$, record a vector of cluster indicator $\xi_X^{(j)}$
Infer a DPM model for the distribution $f_{0,Y}(y)$ using a Gibbs Sampler with N_{it} iterations
$\rightarrow$ for each iteration $1 \leq j \leq N_{\mathrm{it}}$, record a vector of cluster indicator $\xi_Y^{(j)}$

Tests based on contingency tables:
for $1 \leq j \leq N_{\mathrm{it}}$ **do**
  Construct a contingency table $\mathbf{M}^{(j)}$ of size $K_X^{(j)} \times K_Y^{(j)}$ based on $\xi_X^{(j)}$ and $\xi_Y^{(j)}$
  Let $p^{(j)} \leftarrow 1/(1 + BF)$ where BF is defined in (5)
end for

Let $p_{\mathrm{dep}} \leftarrow \frac{1}{n_{\mathrm{it}}} \sum_{j=1}^{n_{\mathrm{it}}} p^{(j)}$

3.2. Mixture model predictive approach

In this section we consider an alternative approach for testing between hypothesis (1). Motivated by Kamary et al. (2014) we replace the testing problem with an estimation one by defining a predictive ensemble model $\mathcal{M}^*$ whose components are the competing models $\mathcal{M}_0$ and $\mathcal{M}_1$. To be precise, let f_0 and f_1 denote the densities of (X, Y) defined by models $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. Then we define a predictive mixture model as a linear combination of sub-models of the form

$$f^*(x, y) = \pi f_1(x, y) + (1 - \pi) f_0(x, y), \tag{10}$$

where π is a free regression parameter with constraint $0 \leq \pi \leq 1$ and $f_0(x, y) = f_{0,X}(x) f_{0,Y}(y)$. This model embeds both $\mathcal{M}_0$ and $\mathcal{M}_1$ for values of π equal to 0 or 1. The main idea of this method is to estimate from the data the mixture parameter π, which indicates the preference of the data for dependence model $\mathcal{M}_1$. In contrast to the latent contingency table procedure this approach requires the explicit construction of a joint model under hypothesis $\mathcal{M}_1$.

Since f_0 and f_1 are unknown densities, we assume Bayesian nonparametric prior distributions. For $f_{0,X}(x)$ and $f_{0,Y}(y)$ we consider the DPM model defined by equations (7) and (8). For f_1 we take a bivariate DPM model defined as

$$f_1(x, y) \sim \int \mathrm{N}(x, y \mid \theta_{X,Y}) G_{X,Y}(\mathrm{d}\theta_{X,Y}), \tag{11}$$

where $\theta_{X,Y} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$G_{X,Y} \sim \mathcal{DP}(c_1, G_1) \tag{12}$$

and $G_1 = \mathrm{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (1/k_0)\boldsymbol{\Sigma})\,\mathrm{IW}(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Psi})$. The parameter π has also to be estimated so we take a prior of the form $\pi \sim \mathrm{Be}(a_0, b_0)$. We ensure that the

centring measures G_0 and G_1 are comparable by setting their hyper-parameters as follows: we have $G_{d-1} = \mathrm{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (1/k_0)\boldsymbol{\Sigma})\,\mathrm{IW}(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Psi})$ for $d = 1$ and 2 with $\nu = d + 2$, the d-dimensional vector $\boldsymbol{\mu}_0 \sim \mathrm{N}(0_d, c_\mu\ \mathbf{I}_d)$, the $d \times d$-matrix $\boldsymbol{\Psi} \sim \mathrm{IW}(\nu, c_\Psi\ \mathbf{I}_d)$ where $\mathbf{I}_d$ is the identity matrix of dimension d. The hyper-parameters c_μ, c_Ψ and k_0 are set to be equal for G_0 and G_1.

Our objective is to highlight pairwise dependence across many pairs of variables, and order the pairs into those showing evidence from strongest to weakest association. This motivates us to consider a simplified method by assessing the relative posterior predictive evidence under $\mathcal{M}_0$ to that of $\mathcal{M}_1$, by calculating an ensemble model using the posterior predictive probability of the observed data $f_1(x_{new}, y_{new}|D)$ and $f_0(x_{new}, y_{new}|D)$ separately. In the following we will use the notations $\hat{f}_j(x_{new}, y_{new}) = f_j(x_{new}, y_{new}|D)$, $j = 0, 1$ to denote the posterior predictive distribution. It is important to note that for all $[p \times (p-1)/2]$ X, Y pairs we use the same prior, and hence same model complexity across all pairs, so ranking by the improvement in posterior predictive likelihood under $\mathcal{M}_1$ relative to $\mathcal{M}_0$ should not *a priori* favour certain pairs over others. This procedure significantly simplifies the inference as we can infer the posterior models by first fitting the three DPM models separately each using the entire sample data, and then updating the ensemble parameter π from its posterior conditional distribution

$$f(\pi \mid \mathrm{D}) \propto f(\pi) \prod_i \left(\pi \hat{f}_1(x_i, y_i) + (1-\pi)\hat{f}_0(x_i, y_i) \right) ,$$

which is a simple line search on $[0, 1]$. We will refer to this inference procedure as MixMod-ensemble – see Algorithm 2.

Algorithm 2 Independence test MixMod-ensemble

Require: Data $\mathrm{D} = \{x_i, y_i\}_{i=1}^n$; Prior parameters a_0 and b_0; Prior parameters for the DPMs
Ensure: Estimate of mixture parameter π

DPMs inference:
$\hat{f}_{0,X} \leftarrow$ posterior prediction of a DPM for distribution of $\{x_i\}_i$ averaged over all Gibbs sampler iteration
$\hat{f}_{0,Y} \leftarrow$ posterior prediction of a DPM for distribution of $\{y_i\}_i$ averaged over all Gibbs sampler iteration
$\hat{f}_1 \leftarrow$ posterior prediction of a DPM for distribution of $\{x_i, y_i\}_i$ averaged over all Gibbs sampler iteration

Estimation of $\hat{\pi}$:
Define a fine grid of $[0, 1]$ with intervals of length $\eta = 10^{-4}$
for $j = 0, \ldots, \eta^{-1}$ **do**
  $\pi^{(j)} \leftarrow j \times \eta$
  $L_j \leftarrow \sum_{i=1}^n \log(\pi^{(j)} \hat{f}_1(x_i, y_i) + (1 - \pi^{(j)})\hat{f}_{0,X}(x_i)\hat{f}_{0,Y}(y_i)) + \log(\mathrm{Be}(\pi^{(j)} \mid a_0, b_0))$
end for
$\hat{\pi} \leftarrow \frac{1}{\sum_j \exp(L_j)} \sum_j \pi^{(j)} \exp(L_j)$

An alternative approach, more closely resembling Kamary et al. (2014), is to consider $\mathcal{M}^*$ as a mixture-model rather than an ensemble model where with probability π the data arises from f_0 and with probability $1-\pi$ from f_1. Diebolt

and Robert (1994) show that posterior sampling in a mixture model is simplified if we introduce latent variable indicators $\zeta_i \sim \text{Ber}(\pi)$ that determine whether observation i comes from f_1, when $\zeta_i = 1$, or from f_0, when $\zeta_i = 0$. Conditional on these latent indicators the mixture components f_0 and f_1 can be updated using only the data points allocated to each model. As noted by Kamary et al. (2014), the Gibbs sampler implemented in this way can become quite inefficient if the parameter π approaches the boundaries $\{0, 1\}$, specially for large sample sizes. We refer to this method as MixMod. For our purposes this requires specifying a Gibbs sampler for the mixture model utilising three DPM models $\{f_1(x, y), f_{0,X}(x), f_{0,Y}(y)\}$ and the mixture allocations for points across all $p \times (p-1)/2$ pairs.

In the paper we will illustrate the performance using MixMod-ensemble, and in the *Supplementary Material* we provide a comparison between MixMod and MixMod-ensemble.

Regardless of the posterior inference procedure, different estimators of π could be obtained from its posterior distribution. We chose to select the expected value as a statistic of dependence, that is,

$$\hat{\pi} = \text{E}(\pi \mid \text{D}) = \int_0^1 \pi \; f(\pi \mid D) d\pi \; . \tag{13}$$

3.3. Computational tractability

Both of the Bayesian non-parametric approaches proposed here are motivated by the increasing necessity of screening large data sets for possible pairwise dependencies where calculation of the formal Bayes factor under $\mathcal{M}_0$ and $\mathcal{M}_1$ is unfeasible or undesirable. In this section, we discuss some computational advantages of our two methods including their amenity to implementation on modern computing architectures exploiting parallelisation on multi-core standalone machines, or clusters of multi-core and many-core machines, or cloud based computing environments.

In relation to parallelisation we see that both methods are divided in two steps: one starts by inferring DPMs using a Gibbs sampler and then perform a dependence test using every iteration of the Gibbs sampler. This decoupling of the inference step and the model comparison step allows to significantly reduce the computational cost of the procedure. In particular, only a couple of thousands of Gibbs sampling iterations are necessary to estimate the predictive posterior densities and posterior distributions over the latent allocation variables. In the environment for statistical computing R (R Core Team, 2014), the parallelisation of both approaches is very simple and only consists in replacing the command *apply* by the command *parLapply* from the package *parallel* – which is included in versions of R following 2.14.0. The R code to run CT-BF and MixMod-ensemble independence tests is available in the *Supplementary Material*.

The CT-BF approach based on the construction of a contingency table is particularly attractive as it is trivially parallelizable and does not involve an

explicit DPM model for the joint $f_1(x, y)$ under $\mathcal{M}_1$. With p measurement variables under study, this approach only needs to infer p independent marginal DPMs, recording information from N_{it} Gibbs sampling iterations for each of them independently in parallel. The MCMC output from the p models is then combined and we perform $N_{\text{it}} \times p \times (p-1)/2$ independent tests where following (5) only involves computing ratios of Gamma functions. As an illustration, in the example described in more details in Section 4, for $p = 562$ measurement variables, the first stage of inference on the DPMs take less than 3 minutes on a 48-core machine, and then the resulting 1.5×10^8 pairwise tests of dependence for all pairs of variables are performed in one hour.

In comparison the MixMod-ensemble approach incurs a greater computational overhead as we require bivariate DPMs, $f_1(x, y)$, to be fit for all pairs. In the illustration below the MixMod-ensemble procedure for the 1.5×10^8 pairs takes approximatively 36 hours on the same 48-core machine.

4. Numerical analysis

4.1. World Health Organisation dataset

In this section, we apply the two approaches described in Section 3 to detect dependencies in economic, social and health indicators from the World Health Organisation (WHO). The WHO Statistical Information System (WHOSIS) has recently been incorporated into the Global Health Observatory (GHO) that contains a data repository (http://www.who.int/gho/database/en/) with mortality and global health estimates, demographic and socioeconomic statistics as well as information regarding health service coverage and risk factors for 194 countries. We combined these datasets to obtain a set of 562 statistics per country. We aim at highlighting potential dependencies between these indicators. Scatterplots of some of these indicators are represented in Figure 1, where for example we see, unsurprisingly, strong dependencies between indicators such as life expectancy at birth and increased life expectancy at age 60 (Pair E).

We applied both the CT-BF and the MixMod-ensemble test to compute the probability of dependence for all the 157,641 pairs of indicators. The two proposed methods require the specification of several parameters of the prior distributions. The impact of these choices is discussed in *Supplementary Material*. For the approach based on contingency tables the prior specifications for models (7) and (8) are set as follows: $c_0 = 10$, $\mu_0 \sim \text{N}(0, 1)$, $k_0 \sim \text{Ga}(1/2, 100/2)$, $\nu = 3$ and $\psi \sim \text{IGa}(1/2, 5)$. Note that c_0 controls the number of clusters induced, so in order to avoid having partitions with only one cluster we set this parameter at a relative large value. To specify the Dirichlet prior for the cell probabilities in the contingency table we took $\alpha_{kl} = 1/2$, which is the Jeffreys prior in a multinomial model. In experimentation we found that the contingency table can be sensitive to the choice of the parameter c_0. This parameter influences the number of clusters in the DPM model and therefore the size of the contingency tables and it is important to specify a value that induces a reasonable number

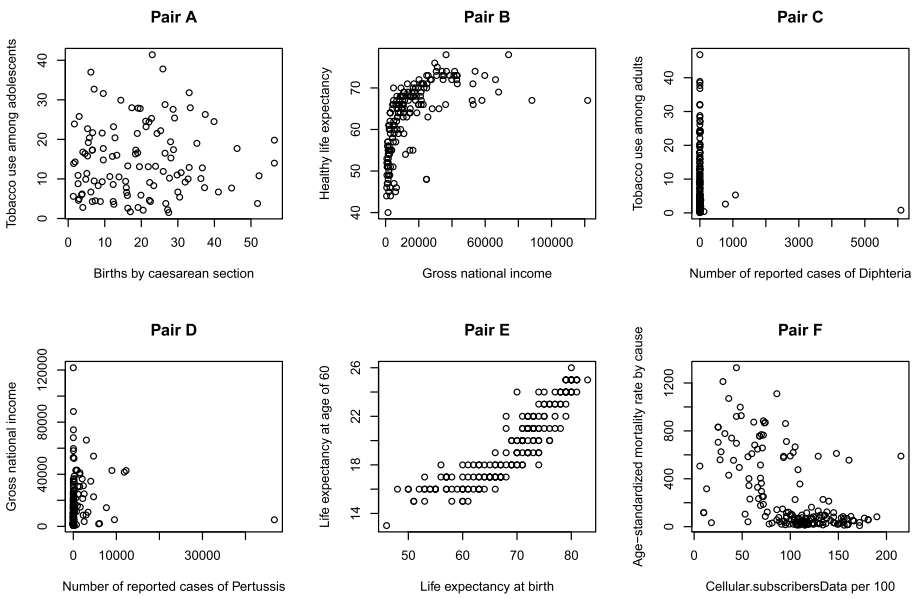

FIG 1. *Examples of the relationship between economic, social and health indicators provided by the WHO Statistical Information System. Each dot corresponds to one country.*

of clusters. We would recommend exploring several values. Results seem fairly insensitive to the choice of the parameters α_{kl} in the Dirichlet priors.

For the approach considering an ensemble mixture model, the parameters c_0 and c_1 are not fixed but specified by $c_0, c_1 \sim \text{Ga}(1,1)$ and $\mu_0 \sim \text{N}(0, 100)$. This change was introduced to allow the model to determine the best fit without constraining the number of clusters. In addition, the prior processes G_0 and G_1 are defined as follows: $G_{d-1} = \text{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (1/k_0)\boldsymbol{\Sigma})\,\text{IW}(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Psi})$ for $d = 1$ and 2 with $\nu = d + 2$, the d-dimensional vector $\boldsymbol{\mu}_0 \sim \text{N}(0_d, 100\,\mathbf{I}_d)$, the $d \times d$-matrix $\boldsymbol{\Psi} \sim \text{IW}(\nu, 0.1\,\mathbf{I}_d)$ and $k_0 \sim \text{Ga}(1/2, 50)$, where $\mathbf{I}_d$ is the identity matrix of dimension d. The prior distribution of the mixing proportion π was specified by taking $a_0 = b_0 = 1/2$. Our experience is that results are fairly robust to the prior parameter settings (see *Supplementary Material*).

The procedures were implemented in the environment for statistical computing R (R Core Team, 2014) and make use of the package *DPpackage* (Jara et al., 2011). Chains were run for 10,000 iterations with a burn in of 1,000 keeping one of every 5th draws for computing estimates.

For both approaches the tests were performed only for pairs containing measurements for at least 10 countries. For the CT-BF approach, the 562 DPMs are inferred using all the available data; however, the contingency tables were constructed taking into account only the countries for which both indicators (in the pair) are available. For the MixMod-ensemble approach, in order to avoid any bias towards one of the two models $\mathcal{M}_0$ or $\mathcal{M}_1$, both the DPMs on the marginals and the DPM on the joint space are inferred only on the countries for

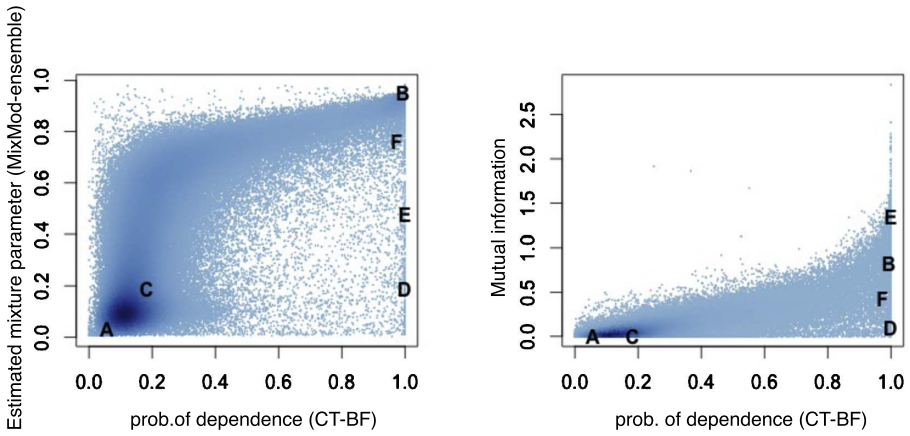

FIG 2. *Performance comparison between the CT-BF and the MixMod-ensemble approaches (left) and the mutual information (right) for every pair of indicators in the WHO dataset. The measures of dependences obtained following CT-BF and MixMod-ensemble are respectively* p_{dep} *and* $\hat{\pi}$*, defined equations* (9) *and* (13) *and approximated following algorithms 1 and 2. The letters A to F correspond to the 6 pairs of indicators illustrated in Figure 1.*

which measurements are available for both indicators. Extending the method to handle missing data is a future objective.

The measure of dependences obtained following our two approaches, i.e. p_{dep} for CT-BF and $\hat{\pi}$ for MixMod-ensemble, defined respectively equations (9) and (13), are compared for each pair of variables in Figure 2 (left panel). Strong dependences (defined as $p_{\text{dep}} > 0.8$) are detected for 5% of pairs, and credible independence (i.e. $p_{\text{dep}} < 0.2$) between 30% of the indicators. We observe that the two probabilistic measures of dependence generally agree for most of the pairs, with the value obtained following the MixMod-ensemble method being generally higher than the probability measure obtained following the CT-BF approach. This elevation in the evidence in dependence is perhaps to be expected as MixMod-ensemble uses the conditional posterior predictive likelihood which will favour the more complex joint model of $f_1(x, y)$. However, the two methods disagree (defined as the probability value obtained following one method is lower than 0.2 while it is larger than 0.8 following the other method) for less than 0.36% of the pairs; and these differences mainly occur when one of the (X, Y) variables is equal to 0 for more than 20% of the countries (see for example pair D).

On balance we prefer to use the CT-BF approach due to its computational scalability, 1 hour of run-time on a 48-core computer in comparison with 36 hours for MixMod-ensemble in this example. We compared the analysis from the CT-BF to that using a mutual information approach computed using the 20-nearest neighbours method, as in Kinney and Atwal (2014) (see Figure 2 right panel where the labelled points correspond to plots in Figure 1). We remark that some pairs of variables with strong dependences under CT-BF have a wide

spread of mutual information, in particular we note pairs D and F that have a probability of dependence close to 1 for CT-BF but relatively low MI values. Visually at least one could argue that associations of the form seen in Figure 1 D and F may be of potential interest to follow up by the analyst.

4.2. Simulation Study for frequentist power analysis

In this section we perform a simulation study to examine the frequentist performance of the two proposed tests on some controlled scenarios. The objective is to verify that we are not losing much power against a popular non-probabilistic method based on mutual information, which is optimised for frequentist power. Simulated datasets are generated under the following four different scenarios:

1. A bivariate normal model: $(X, Y) \sim \mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$,
2. A sinusoidal model: $Y = 2\sin(X) + \eta$, with $\eta \sim \mathrm{N}(0, \phi^2)$, and $X \sim \mathrm{Un}[0, 5\pi]$
3. A parabolic model: $Y = 2X^2/3 + \eta$, with $\eta \sim \mathrm{N}(0, \phi^2)$, and $X \sim \mathrm{N}(0, 1)$
4. A circular model: $X = 10\cos(\theta) + \eta$ and $Y = 10\sin(\theta) + \eta$, with $\theta \sim \mathrm{Un}[0, 2\pi]$ and $\eta \sim \mathrm{N}(0, \phi^2)$.

For the sinusoidal, parabolic and circular models, the parameter ϕ controls the level of noise, whereas for the normal model the correlation ρ controls the degree of dependence between the two samples. We generated fifty independent datasets from each model with a sample size $n = 250$ with different correlations $\rho \in \{0, 0.1, 0.3, 0.5, 0.9\}$, for model (a), and levels of noise $\phi \in \{1, 2, 3, 4, 5\}$ for models (b)–(d). Figure 3 shows one of the fifty simulated dataset as illustration.

For all the simulated datasets we apply our different procedures for testing hypothesis (1). We use the same priors specifications as described in Section 4.1.

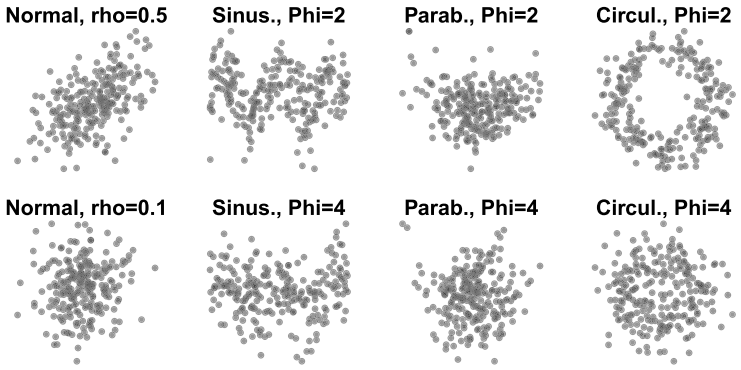

FIG 3. *Samples of size* 250 *generated from the four scenarios for two levels of correlation* ρ *in the normal model and two levels of noise* ϕ *in the sinusoidal, parabolic and circular models.*

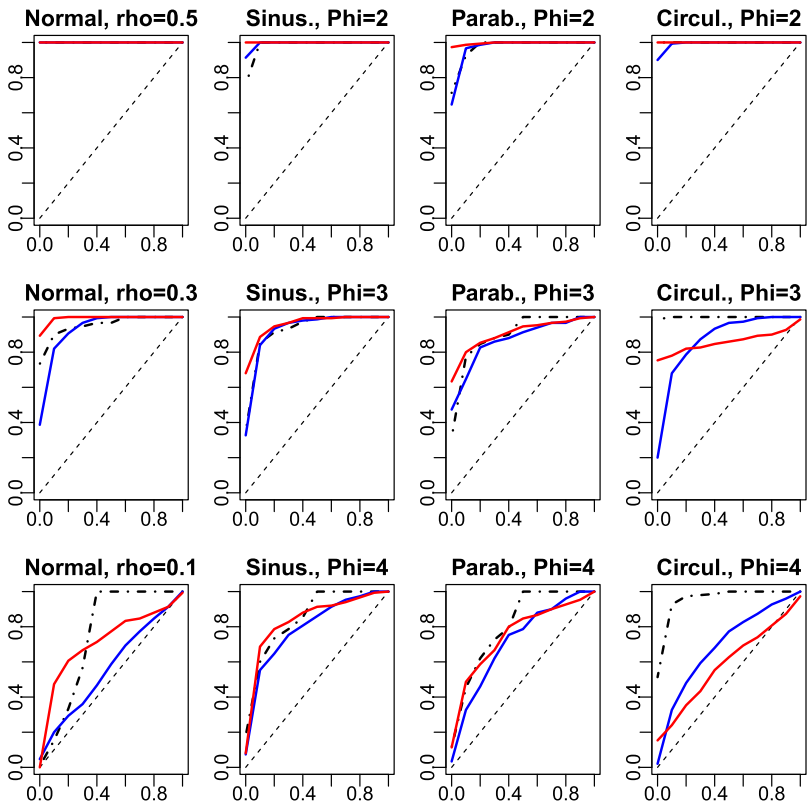

FIG 4. *ROC curves for competing methods as a function of correlation and noise level for models (a)–(d). CT-BF (blue line), MixMod-ensemble (red line) and Mutual Information approximated using the* 20 *nearest neighbours (black dotted line).*

To investigate the power of the two approaches, we create ROC curves that compare the rate of true positives (percentage of times the procedure detects dependence among the fifty datasets generated from a dependent model) and false positives (percentage of times the procedure detects dependence among fifty null datasets generated by randomly permuted the indexes of the two samples to destroy any dependences) for different threshold values. We also compare the performance of the proposed methods to the current state of the art conventional method, which is based on mutual information (using the 20 nearest neighbours). The ROC curves are reported in Figure 4; see also *Supplementary Material* that contains additional more extensive comparisons.

We observe that the proposed methods have similar performances to the current leading conventional method for data coming from a sinusoidal or a parabolic model. For data generated from the circular model however the mutual information method outperforms our approaches.

5. Conclusion

We presented two Bayesian nonparametric procedures for highlighting pairwise dependencies between random variables that are scalable to large data sets. The methods make use of standard software in R for implementing DPM of Gaussians and are designed to exploit modern computer architectures. As such they are readily amenable to applied statisticians interested in exploratory analysis of large data sets. A power analysis shows that the procedures are comparable with that of current non-Bayesian methods based on mutual information, while having the advantage of being probabilistic in their measurement.

Acknowledgements

This work was done whilst Luis Nieto-Barajas was visiting the Department of Statistics at the University of Oxford. He was supported by CONACYT grant 244459 and *Asociacion Mexicana de Cultura, A.C.* Mexico. Chris Holmes is supported by the Oxford-Man Institute, the Alan Turing Institute, the Engineering and Physical Sciences Research Council programme grant i-like EP/K014463/1, and the Medical Research Council, UK, grant MC_UP_A390_1107.

Supplementary Material

Supplement to "Scalable Bayesian nonparametric measures for exploring pairwise dependence via Dirichlet Process Mixtures"
(doi: 10.1214/16-EJS1171SUPP; .pdf).

References

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

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. MR1122806

D. B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218, 2006.

J. Diebolt and C. Robert. Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, pages 363–375, 1994.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973. MR0350949

T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Academic Press, New York, 1983. MR0736538

S. Filippi and C. C. Holmes. A Bayesian nonparametric approach to quantifying dependence between random variables. *Bayesian Analysis*, doi: 10.1214/16-BA1027.

I. J. Good and J. F. Crook. The robustness and sensitivity of the mixed Dirichlet Bayesian test for independence in contingency tables. *Annals of Statistics*, 15:670–693, 1987.

E. Gunel and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61:545–557, 1974.

C. C. Holmes, F. Caron, J. E. Griffin, D. A. Stephens, et al. Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, 10(2):297–320, 2015. MR3420884

A. Jara, T. Hanson, F. Quintana, P. Mueller, and G. Rosner. DPpackage: Bayesian semi- and nonparametric modeling in r. *Journal of Statistical Software*, pages 1–30, 2011.

K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. Testing hypotheses via a mixture estimation model. *Preprint arXiv:1412.2044*, 2014.

J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014. MR3200177

A. Y. Lo et al. On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984. MR0733519

E. F. Lock and D. B. Dunson. Two-sample testing with dirichlet mixtures. *Preprint arXiv:1311.0307*, 2013.

K. Pearson. On the χ^2 test of goodness of fit. *Biometrika*, 14:186–191, 1922.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

A. F. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.