entropy

MDPI

Article

Bayesian Dependence Tests for Continuous, Binary and Mixed Continuous-Binary Variables

Alessio Benavoli [1,*] and Cassio P. de Campos [2]

[1] Dalle Molle Institute for Artificial Intelligence, Manno 6928, Switzerland
[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK; c.decampos@qub.ac.uk
* Correspondence: alessio@idsia.ch; Tel.: +41-58-666-6509

Academic Editors: Julio Stern and Adriano Polpo
Received: 20 June 2016; Accepted: 26 August 2016; Published: 6 September 2016

Abstract: Tests for dependence of continuous, discrete and mixed continuous-discrete variables are ubiquitous in science. The goal of this paper is to derive Bayesian alternatives to frequentist null hypothesis significance tests for dependence. In particular, we will present three Bayesian tests for dependence of binary, continuous and mixed variables. These tests are nonparametric and based on the Dirichlet Process, which allows us to use the same prior model for all of them. Therefore, the tests are "consistent" among each other, in the sense that the probabilities that variables are dependent computed with these tests are *commensurable* across the different types of variables being tested. By means of simulations with artificial data, we show the effectiveness of the new tests.

Keywords: dependence; Bayesian independence test; Dirichlet Process

1. Introduction

Tests for dependence of continuous, discrete and mixed continuous-discrete variables are fundamental in science. The standard way to statistically assess if two (or more) variables are dependent is by using null-hypothesis significance tests (NHST), such as χ^2-test, Kendall's τ, etc. However, these tests are affected by the drawbacks which characterize NHST [1–3]. An NHST computes the probability of getting the observed (or a larger) value of the statistics under the assumption that the null hypothesis of independence is true, which is obviously not the same as the probability of variables being dependent on each other, given the observed data. Another common problem is that the claimed statistical significance might have no practical impact. Indeed, the usage of NHST often relies on the wrong assumptions that p-values are a reasonable proxy to the probability of the null hypothesis and that statistical significance implies practical significance.

In this paper, we propose a *collection of Bayesian dependence tests*. The questions we are actually interested in—for example, *Is variable Y dependent on Z?* or *Based on the experiments, how probable is Y dependent on Z?*—are actually questions about posterior probabilities. Answers to these questions are naturally provided by Bayesian methods. The core of this paper is thus to derive Bayesian alternatives to frequentist NHST and to discuss their inference and results. In particular, we present three Bayesian tests for dependence of binary, continuous and mixed variables. All of these tests are nonparametric and based on the Dirichlet Process. This allows us to use the same prior model for all the tests we develop. Therefore, they are "consistent" in the sense that the probabilities of dependence we compute are *commensurable* across the tests. This is another main difference about such an approach and the use of p-values, since the latter usually cannot be compared across different types of tests.

To address the issue of how to choose the prior parameters in case of lack of information, we propose the use of the Imprecise Dirichlet Process (IDP) [4]. It consists of a family of Dirichlet

processes with fixed prior strength and and prior probability measure free to span the set of all distributions. In this way, we obtain as a byproduct a measure of sensitivity of inferences to the choice of the prior parameters.

Nonparametric tests based on the Dirichlet Process and on similar ideas to those presented in this paper have also been proposed in [4] to develop a Bayesian rank test, in [5] for a Bayesian signed-rank test, in [6] for a Bayesian Friedman test and in [7] for a Bayesian test that accounts for censored data.

Several alternative Bayesian methods are available for testing of independence. The test of linear dependence between two continuous univariate random variables can be achieved by fitting a linear model and inspecting the posterior distribution of the correlation coefficient. A more sophisticated test based on a Dirichlet Process Mixture prior is instead presented in [8] to deal with linear and nonlinear dependences. Other methods were proposed for testing of independence based on a contingency table [9–11]. The main difference between these works and the work presented in this paper is that we provide tests for continuous, categorical (binary) and mixed variables using the same approach. This allows us to derive a very general framework to test independence/dependence (these tests could be used for instance for feature selection in machine learning [12–15]).

By means of simulations on artificial data, we use our test to decide if two variables are dependent. We show that our Bayesian test achieves equal or better results than the frequentist tests. We moreover show that the IDP test is more robust, in the sense that it acknowledges when the decision is *prior-dependent*. In other words, the IDP test suspends the judgment and becomes *indeterminate* when the decision becomes prior dependent. Since IDP has all the positive features of a Bayesian test and it is more reliable than the frequentist tests, we propose IDP as a new test for testing dependence.

2. Dirichlet Process

The Dirichlet Process was developed by Ferguson [16] as a probability distribution on the space of probability distributions. Let $\mathbb{X}$ be a standard Borel space with Borel σ-field $\mathcal{B}_{\mathbb{X}}$ and $\mathbb{P}$ be the space of probability measures on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$ equipped with the weak topology and the corresponding Borel σ-field $\mathcal{B}_{\mathbb{P}}$. Let $\mathbb{M}$ be the class of all probability measures on $(\mathbb{P}, \mathcal{B}_{\mathbb{P}})$. We call the elements $\mu \in \mathbb{M}$ nonparametric priors.

An element of $\mathbb{M}$ is called a Dirichlet Process distribution $\mathcal{D}(\alpha)$ with base measure α if for every finite measurable partition $B_1, \ldots, B_m$ of $\mathbb{X}$, the vector $(P(B_1), \ldots, P(B_m))$ has a Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_m))$, where $\alpha(\cdot)$ is a finite positive Borel measure on $\mathbb{X}$. Consider the partition $B_1 = A$ and $B_2 = A^c = \mathbb{X} \backslash A$ for some measurable set $A \in \mathbb{X}$, then if $P \sim \mathcal{D}(\alpha)$ from the definition of the DP we have that $(P(A), P(A^c)) \sim Dir(\alpha(A), \alpha(\mathbb{X}) - \alpha(A))$, which is a β distribution. From the moments of the β distribution, we can thus derive that:

$$\mathcal{E}[P(A)] = \frac{\alpha(A)}{\alpha(\mathbb{X})}, \quad \mathcal{E}[(P(A) - \mathcal{E}[P(A)])^2] = \frac{\alpha(A)(\alpha(\mathbb{X}) - \alpha(A))}{(\alpha(\mathbb{X})^2(\alpha(\mathbb{X}) + 1))}, \tag{1}$$

where we have used the calligraphic letter $\mathcal{E}$ to denote expectation with respect to the Dirichlet process. This shows that the normalized measure $\alpha(\cdot)/\alpha(\mathbb{X})$ of the DP reflects the prior expectation of P, while the scaling parameter $\alpha(\mathbb{X})$ controls how much P is allowed to deviate from its mean $\alpha(\cdot)/\alpha(\mathbb{X})$. Let $s = \alpha(\mathbb{X})$ stand for the total mass of $\alpha(\cdot)$ and $\alpha^*(\cdot) = \alpha(\cdot)/s$ stand for the probability measure obtained by normalizing $\alpha(\cdot)$. If $P \sim \mathcal{D}(\alpha)$, we shall also describe this by saying $P \sim Dp(s, \alpha^*)$ or, if $\mathbb{X} = \mathbb{R}$, $P \sim Dp(s, G_0)$, where G_0 stands for the cumulative distribution function of α^*.

Let $P \sim Dp(s, \alpha^*)$ and f be a real-valued bounded function defined on $(\mathbb{X}, \mathcal{B})$. Then the expectation with respect to the Dirichlet Process of $E[f]$ is

$$\mathcal{E}\left[E(f)\right] = \mathcal{E}\left[\int f dP\right] = \int f d\mathcal{E}[P] = \int f d\alpha^*. \tag{2}$$

One of the most remarkable properties of the DP priors is that the posterior distribution of P is again a DP. Let $X_1, \ldots, X_n$ be an independent and identically distributed sample from P and $P \sim Dp(s, \alpha^*)$, then the posterior distribution of P given the observations, denoted as $P_{X|X^n}$, is

$$P_{X|X^n} \sim Dp\left(s+n, \frac{s}{s+n}\alpha^* + \frac{1}{s+n}\sum_{i=1}^{n}\delta_{X_i}\right), \tag{3}$$

where δ_{X_i} is an atomic probability measure centered at X_i and $X^n = \{X_1, \ldots, X_n\}$. This means that the Dirichlet Process satisfies a property of conjugacy, in the sense that the posterior for P is again a Dirichlet Process with updated unnormalized base measure $\alpha + \sum_{i=1}^{n}\delta_{X_i}$. From Equations (1)–(3), we can easily derive the posterior mean and variance of $P(A)$ and, respectively, posterior expectation of f. Hereafter we list some useful properties of the DP that will be used in the sequel (see Chapter 3 in [17]).

(a) In case $\mathbb{X} = \mathbb{R}$, since P is completely defined by its cumulative distribution function F, a-priori we say $F \sim Dp(s, G_0)$ and a posteriori we can rewrite (3) as follows:

$$F_{X|X^n} \sim Dp\left(s+n, \frac{s}{s+n}G_0 + \frac{n}{s+n}\frac{1}{n}\sum_{i=1}^{n}I_{[X_i,\infty)}\right), \tag{4}$$

where I is the indicator function and $\frac{1}{n}\sum_{i=1}^{n}I_{[X_i,\infty)}$ is the empirical cumulative distribution.

(b) Consider an element $\mu \in \mathbb{M}$ which puts all its mass at the probability measure $P = \delta_x$ for some $x \in \mathbb{X}$. This can also be modeled as $Dp(s, \delta_x)$ for each $s > 0$.

(c) Assume that $P_1 \sim Dp(s_1, \alpha_1^*)$, $P_2 \sim Dp(s_2, \alpha_2^*)$, $(\omega_1, \omega_2) \sim Dir(s_1, s_2)$ and P_1, P_2, (ω_1, ω_2) are independent, then Section 3.1.1. in [17]:

$$\omega_1 P_1 + \omega_2 P_2 \sim Dp\left(s_1+s_2, \frac{s_1}{s_1+s_2}\alpha_1^* + \frac{s_2}{s_1+s_2}\alpha_2^*\right). \tag{5}$$

(d) Let $P_{X|X^n}$ have distribution $Dp(s+n, \frac{s}{s+n}\alpha^* + \frac{1}{s+n}\sum_{i=1}^{n}\delta_{X_i})$. We can write

$$P_{X|X^n} = \omega_0 P + \sum_{i=1}^{n}\omega_i \delta_{X_i}, \tag{6}$$

where $\sum_{i=0}^{n}\omega_i = 1$, $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$ and $P \sim Dp(s, \alpha^*)$. This follows from (b)–(c).

An issue in the use of the DP as prior measure on P is how to choose the infinite dimensional parameter G_0 in case of lack of prior information. There are two avenues that we can follow. The first assumes that prior ignorance can be modelled satisfactorily by a so-called noninformative prior. In the DP setting, the only noninformative prior that has been proposed so far is the limiting DP obtained for $s \to 0$, which has been introduced by [16] and discussed by [18]. The second approach suggests that lack of prior information should be expressed in terms of a set of probability distributions. This approach known as *Imprecise Probability* [19–21] is connected to *Bayesian robustness* [22–24] and it has been extensively applied to model prior (near-)ignorance in parametric models. In this paper, we implement a prior (near-)ignorance model by considering a set of DPs obtained by fixing s to a strictly positive value and letting G_0 span the set of all distributions. This model has been introduced in [4] with the name of Imprecise Dirichlet Process (IDP).

3. Bayesian Independence Tests

Let us denote by X the vector of variables $[Y, Z]^T$ so that the n observations of X can be rewritten as

$$X^n = \{X_1, \ldots, X_n\}, \tag{7}$$

that is, a set of n vector-valued i.i.d. observations of X. We also consider an auxiliary variable X' together with X. We assume that X, X' are independent variables from the same unknown distribution and that $X'^n = X^n$, that is, we have the same observations of X and X'.

Let P be the unknown distribution of X, X' and assume that the prior distribution of P is $Dp(s, \alpha^*)$. Our goal is to compute the posterior of P. The posterior of P is given in (3) and, by exploiting (6), we know that

$$P_{X|X^n} \sim \omega_0 P + \sum_{i=1}^{n} \omega_i \delta_{X_i}, \tag{8}$$

with $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$ and $P \sim Dp(s, \alpha^*)$. The distribution $P_{X'|X'^n}$ of X' is similarly defined.

The questions we pose in a statistical analysis can all be answered by querying this posterior distribution in different ways. We adopt this posterior distribution to devise Bayesian counterparts of the independence hypothesis tests.

3.1. Bayesian Bivariate Independence Test for Binary Variables

Let us assume that the variables $Y, Z \in \{0, 1\}$ (that is, they are binary). Our aim is to devise a Bayesian independence test for binary variables based on the DP. We will also show that our test is a Bayesian generalisation of the frequentist χ^2-test for independence applied to binary variables. We start by defining the following quantities:

$$E\left(I_{\{0,0\}}(X) I_{\{1,1\}}(X') | X^n, X'^n\right) = \iint I_{\{0,0\}}(X) I_{\{1,1\}}(X') dF(X|X^n) dF(X'|X'^n),$$

where we have exploited the independence of X, X' and here $F(X|X^n)$ denotes the posterior cumulative distribution of $P_{X|X^n}$ defined in (8). From (8), it can easily be verified that

$$E\left(I_{\{0,0\}}(X) I_{\{1,1\}}(X') | X^n, X'^n\right) = \omega_{00}\omega_{11},$$

where

$$\omega_{00} = \omega_0 \int I_{\{0,0\}}(X) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{0,0\}}(Y_i, Z_i), \tag{9}$$

and

$$\omega_{11} = \omega_0 \int I_{\{1,1\}}(X) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{1,1\}}(Y_i, Z_i), \tag{10}$$

where in the last equality we have exploited the fact that X' has the same distribution as X and also the same observations. The two quantities ω_{00}, ω_{11} include two terms. The first is the term due to the prior $dF \sim Dp(s, \alpha^*)$ and the second term is due to the observations.

Similarly, we compute

$$E\left(I_{\{0,1\}}(X) I_{\{1,0\}}(X') | X^n, X'^n\right) = \omega_{01}\omega_{10},$$

where

$$\omega_{01} = \omega_0 \int I_{\{0,1\}}(X) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{0,1\}}(Y_i, Z_i), \tag{11}$$

and

$$\omega_{11} = \omega_0 \int I_{\{1,0\}}(X) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{1,0\}}(Y_i, Z_i). \tag{12}$$

Summing up, $\omega_{00}, \omega_{1,0}, \omega_{0,1}, \omega_{11}$ represent the posterior probabilities of the events $(0,0)$ (that is, $Y = 0$ and $Z = 0$), $(1,0)$, $(0,1)$ and $(1,1)$, respectively, according to the posterior joint distribution $F(X|X^n)$.

Theorem 1. *The variables Y and Z are said to be concordant (dependent) with posterior probability $(1-\gamma)$ provided that*

$$\mathcal{P}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0 | X^n) > (1-\gamma), \tag{13}$$

and they are said to be discordant provided that

$$\mathcal{P}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) < 0 | X^n) > (1-\gamma), \tag{14}$$

where $\mathcal{P}$ is the probability computed with respect to $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$ and $dF \sim Dp(s, \alpha^*)$. Finally, they are said to be simply dependent with posterior probability $(1-\gamma)$ provided that

$$0 \notin (1-\gamma) HDI(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) | X^n), \tag{15}$$

where HDI denotes the posterior Highest Density Interval of $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$.

Proof. We just derive the third statement. The other two statements are analogue. We first consider the indicator functions

$$I_{\{0\}}(Y),\ I_{\{1\}}(Y),\ I_{\{0\}}(Z),\ I_{\{1\}}(Z), \tag{16}$$

and same for the auxiliary variables Y', Z'. By computing the expectation of these functions, we can obtain the *marginals* of the variables Y, Z with respect to the joint P_X:

$$\omega_{0\bullet} := E(I_{\{0\}}(Y)|X^n) = \omega_0 \int I_{\{0\}}(Y) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{0\}}(Y_i), \tag{17}$$

$$\omega_{1\bullet} := E(I_{\{1\}}(Y)|X^n) = \omega_0 \int I_{\{1\}}(Y) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{1\}}(Y_i), \tag{18}$$

$$\omega_{\bullet 0} := E(I_{\{0\}}(Z)|X^n) = \omega_0 \int I_{\{0\}}(Z) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{0\}}(Z_i), \tag{19}$$

$$\omega_{\bullet 1} := E(I_{\{1\}}(Z)|X^n) = \omega_0 \int I_{\{1\}}(Z) dF(X) + \sum_{i=1}^{n} \omega_i I_{\{1\}}(Z_i), \tag{20}$$

where $\omega_{0\bullet}$ (resp. $\omega_{1\bullet}$) denotes the marginal with respect to Z when $Y = 0$ (resp. $Y = 1$), while $\omega_{\bullet 0}$ (resp. $\omega_{\bullet 1}$) denotes the marginal with respect to Y when $Y = 0$ (resp. $Y = 1$).

Then, by exploiting independence between X and X', we derive

$$E(I_{\{0\}}(Y)I_{\{0\}}(Z')|X^n, X'^n) = \omega_{0\bullet}\omega_{\bullet 0},\ \ E(I_{\{1\}}(Y)I_{\{0\}}(Z')|X^n, X'^n) = \omega_{1\bullet}\omega_{\bullet 0}, \tag{21}$$

$$E(I_{\{0\}}(Y)I_{\{1\}}(Z')|X^n, X'^n) = \omega_{0\bullet}\omega_{\bullet 1},\ \ E(I_{\{1\}}(Y)I_{\{1\}}(Z')|X^n, X'^n) = \omega_{1\bullet}\omega_{\bullet 1}. \tag{22}$$

We are now ready to define the independence test. If the two variables Y, Z are independent, then the vector

$$v = (\omega_{00}, \omega_{10}, \omega_{01}, \omega_{11}) - (\omega_{0\bullet}\omega_{\bullet 0}, \omega_{1\bullet}\omega_{\bullet 0}, \omega_{0\bullet}\omega_{\bullet 1}, \omega_{1\bullet}\omega_{\bullet 1}),$$

has zero mean. Note that the first component of the vector v is $E(I_{\{0,0\}}(X) - I_{\{0\}}(Y)I_{\{0\}}(Z')|X^n, X'^n)$ and thus is a well-defined quantity with respect to our probabilistic model (similarly for the other terms). Therefore, the independence test reduces to checking whether the $(1-\gamma)\%$ highest density credible region (HCR) of v includes the zero vector. It can be easily verified that $|v_l| = |v_m|$ for each l, m component of v. In fact, we have

$$\omega_{i\bullet} = \omega_{ij} + \omega_{i\bar{j}}, \;\; \omega_{\bullet j} = \omega_{ij} + \omega_{\bar{i}j},$$

for $i, j \in \{0,1\}$ and $\bar{i} = 1 - i$, $\bar{j} = 1 - j$, and so

$$\begin{aligned} \omega_{ij} - (\omega_{ij} + \omega_{i\bar{j}})(\omega_{ij} + \omega_{\bar{i}j}) &= \omega_{ij}(\omega_{ij} + \omega_{\bar{i}j} + \omega_{i\bar{j}} + \omega_{\bar{i}\bar{j}}) - (\omega_{ij} + \omega_{i\bar{j}})(\omega_{ij} + \omega_{\bar{i}j}) \\ &= \omega_{ij}\omega_{\bar{i}\bar{j}} - \omega_{i\bar{j}}\omega_{\bar{i}j}. \end{aligned}$$

Therefore, it is enough to check whether

$$0 \notin (1-\gamma)\%\; HDI(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})).$$

If this is the case, then we can declare that the two variables are dependent with probability $(1-\gamma)$. Here, the multiplier 2 in $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$ is only a scaling factor so that $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$ varies in $[-0.5, 0.5]$. □

From the proof of Theorem 1 it is evident the similarity of the test with the frequentist χ^2-test for independence. Both tests use the difference between the joint and the product of the marginals as a measure of dependence. The advantage of the Bayesian approach is that we compute posterior probabilities for the hypothesis in which we are interested and not the probability of getting the observed (or a larger) difference under the assumption that the null hypothesis of independence is true.

The probabilities computed in Theorem 1 depend on the prior information $F \sim Dp(s, \alpha^*)$. In this paper we adopt IDP as prior model. We can then perform a Bayesian nonparametric test that is based on extremely weak prior assumptions, and easy to elicit, since it requires only the choice of the strength s of the DP instead of its infinite-dimensional parameter α^*. The infinite-dimensional parameter α^* is free to vary in the set of all distributions.

Let us consider for instance (13). Each one of these priors gives a posterior probability $\mathcal{P}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$. We can characterize this set of posteriors by computing the lower and upper bounds $\underline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$ and $\overline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$. Inferences with IDP can be computed by verifying if

$$\underline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n) > (1-\gamma), \;\; \overline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n) > (1-\gamma), \tag{23}$$

and then by taking the following decisions:

1. if both the inequalities are satisfied, then we declare that the two variables are dependent with probability larger than $1-\gamma$;
2. if only one of the inequalities is satisfied (which has necessarily to be the one for the upper), we are in an indeterminate situation, that is, we cannot decide;
3. if both are not satisfied, then we declare that the probability that the two variables are dependent is lower than the desired probability of $1-\gamma$.

When IDP returns an indeterminate decision, it means that the evidence from the observations is not enough to declare that the probability of the hypothesis being true is either larger or smaller than the desired value $1-\gamma$; more observations are necessary to reach a reliable decision.

Theorem 2. *The upper probability* $\overline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$ *is obtained by a prior measure* $\alpha^* = m\delta_{(0,0)} + (1-m)\delta_{(1,1)}$ *with*

$$m = \begin{cases} 0 & \text{if } e_1 + \omega_0 < e_0\,, \\ 1 & \text{if } e_0 < e_1 - \omega_0\,, \\ \dfrac{\omega_0 + e_1 - e_0}{2\omega_0} & \text{other,} \end{cases} \tag{24}$$

where $e_0 = \sum_{i=1}^n \omega_i I_{\{0,0\}}(Y_i, Z_i)$ *and* $e_1 = \sum_{i=1}^n \omega_i I_{\{1,1\}}(Y_i, Z_i)$. *The lower probability* $\underline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$ *is obtained by a prior measure* $\alpha^* = m\delta_{(1,0)} + (1-m)\delta_{(0,1)}$ *with the same* m *as before but* $e_0 = \sum_{i=1}^n \omega_i I_{\{1,0\}}(Y_i, Z_i)$ *and* $e_1 = \sum_{i=1}^n \omega_i I_{\{0,1\}}(Y_i, Z_i)$.

Proof. We are interested in the quantity $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$. It is clear that in order to maximize the probability that $\omega_{00}\omega_{11} - \omega_{01}\omega_{10} > 0$ we must put all the prior mass on $\omega_{00}\omega_{11}$. Let us call $m = I_{\{0,0\}}(X)dF(X)$. Then $\int I_{\{1,1\}}(X)dF(X) = 1 - I_{\{0,0\}}(X)dF(X) = 1 - m$. From (9)–(12), we have that $\omega_{00} = \omega_0 m + e_0$ and $\omega_{11} = \omega_0(1-m) + e_1$ with $m \in [0,1]$. By computing the derivative with respect to m we have

$$\frac{d}{dm}\omega_{00}\omega_{11} = \omega_0\left(\omega_0(1-m) + e_1\right) + \left(\omega_0 m + e_0\right)\left(-\omega_0\right)\,,$$

whose zero is $m = \dfrac{\omega_0 + e_1 - e_0}{2\omega_0}$, which is also a maximum. Hence, the maximum can be either on $m = \dfrac{\omega_0 + e_1 - e_0}{2\omega_0}$ or on the extremes $m = 0$ or $m = 1$. This can be easily verified by checking when $\omega_0 + e_1 - e_0 < 0$ (so $m = 0$) or $\omega_0 + e_1 - e_0 > 2\omega_0$ (so $m = 1$). The lower probability $\underline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$ can be determined using a similar reasoning. □

Since $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$, the computation of $\underline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$, $\overline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n)$ can be obtained by Monte Carlo sampling. The following pseudo-code describes how to compute the upper (the lower can be computed in a similar way).

1. Initialize the counter P_c to 0 and the array V to empty;
2. For $i = 1, \ldots, N_{mc}$
   (a) sample $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$;
   (b) compute $\omega_{00}, \omega_{01}, \omega_{10}, \omega_{11}$ as in (9)–(12) by choosing $dF(X) = m\delta_{(0,0)}(X) + (1-m)\delta_{(1,1)}(X)$ with m defined in Theorem 2;
   (c) compute $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$ and store the result in V;
   (d) if $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0$ then $P_c = P_c + 1$ else $P_c = P_c + 0$.
3. compute the histogram of the elements in V (this gives us the plot of the posterior of $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$)
4. compute the posterior upper probability that $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$ is greater than zero as $\overline{\mathcal{P}}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}) > 0|X^n) \approx P_c / N_{mc}$.

The number of Monte Carlo samples N_{mc} is equal to 100 thousand in the next examples and figures.

The lower and upper HDI intervals in Theorem 1 can also be obtained as in Theorem 2 and computed via Monte Carlo sampling (HDI can be computed using the values stored in V, see pseudo-code). Hereafter we will denote the two intervals corresponding to the lower and upper distributions as $\underline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$ and $\overline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$, respectively.

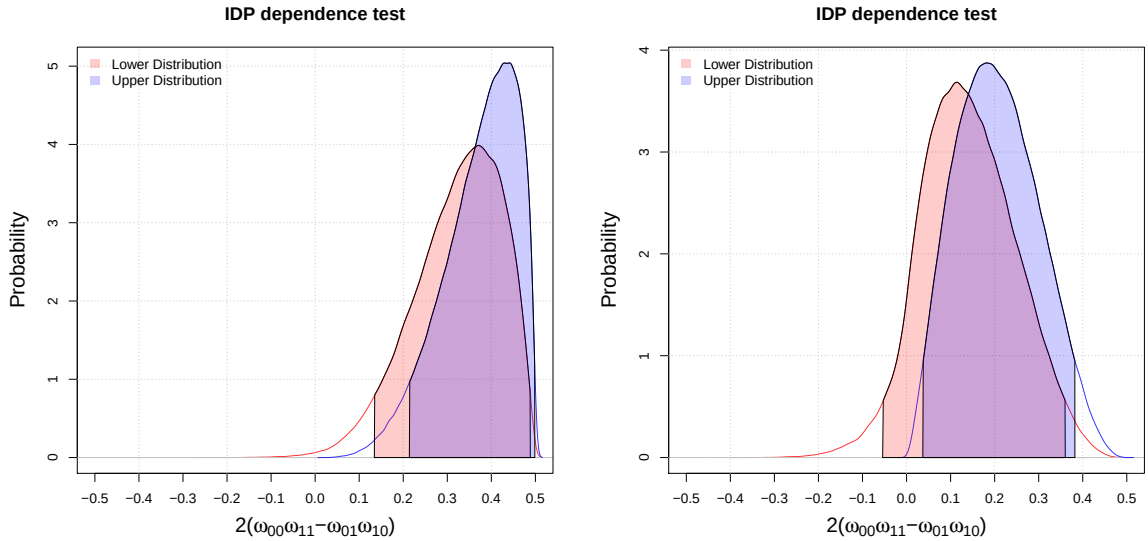
The only prior parameter that must be selected with IDP is the prior strength s. The value of s determines how quickly the posteriors corresponding to the lower and upper probabilities converge as the number of observations increases. We select $s = 0.5$—this means that we need at least 4 concordant binary observations to take a decision with $1 - \gamma = 0.95$. In other words, for $s = 0.5$

we need two observations of type $Y = 0, Z = 0$ and two of type $Y = 1, Z = 1$ to guarantee that both $1 - \gamma = 0.95\%$ HDI intervals, i.e., $\underline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$ and $\overline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$, do not include the zero. For any number of (and configuration of) observations less than four, the test is always indeterminate (i.e., no decision can be taken). Thus, four is the minimum number of observations that is required to take a decision. This choice is arbitrary and subjective, but is our measure of cautiousness. We make clearer the meaning of determinate and indeterminate in the following example.

Example 1. *Let us consider the following three matrices of* 10 *paired binary i.i.d. observations*

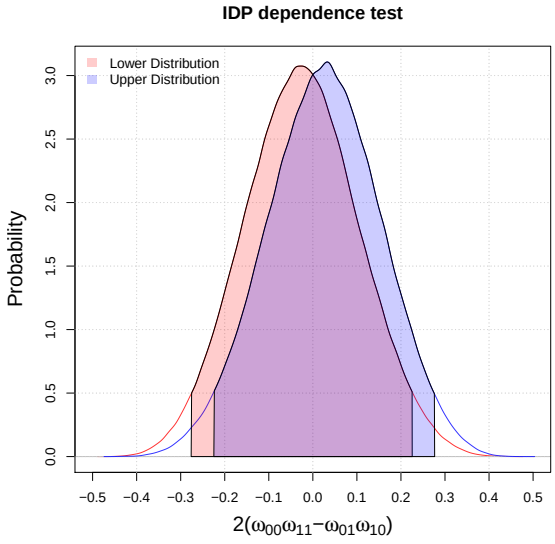
$$X_a^{10} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}^T,$$
$$X_b^{10} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \quad (25)$$
$$X_c^{10} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T.$$

They correspond to different degrees of dependence. Figure 1 shows the lower and upper distributions of $2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10})$ *and the relative* 95% *HDI, i.e.,* $\underline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$ *and* $\overline{HDI}(2(\omega_{00}\omega_{11} - \omega_{01}\omega_{10}))$*, for the three cases* a, b, c *(the filled in areas). In case (a), the two variables are dependent (concordant) with probability greater than* 0.95*, since all the mass of the lower and upper distributions are in the interval* $[0, 0.5]$*. In the second case, we are in an indeterminate situation, that is, the lower and upper are in disagreement, which means that the inference is prior dependent. In the third case, we can only say that they are not dependent at* 95% *since both the HDI intervals include the zero.*



(a) (b)

Figure 1. *Cont.*



(c)

Figure 1. Three possible results of the independence hypothesis testing with two binary variables. The red and blue filled areas correspond respectively to the lower and upper HDI. (**a**) Dependent at 95% ; (**b**) Indeterminate at 95%; (**c**) They are not dependent at 95%.

3.2. Bayesian Bivariate Independence Test for Continuous Variables

Let us assume that variables $Y, Z \in \mathbb{R}$, that is, they are real continuous variables. Our aim is to devise a Bayesian independence test for continuous variables based on the DP. We will also show that our test is a Bayesian generalisation of Kendall-τ test for independence. This test uses results from [25] that derived a Bayesian Kendall's τ statistic using DP. As before, we introduce auxiliary variables Y', Z'. We start by defining the following quantities:

$$\begin{aligned} T_1 &= \{(Y, Z, Y', Z') : (Y - Y')(Z - Z') > 0\}, \\ T_2 &= \{(Y, Z, Y', Z') : (Y - Y')(Z - Z') < 0\}. \end{aligned}$$

T_1 and T_2 are concordance measures. We can then compute

$$E[I_{T_1} - I_{T_2}] = \iint (I_{T_1}(X, X') - I_{T_2}(X, X'))dF(X|X^n)dF(X'|X'^n), \tag{26}$$

where we have exploited the independence of X, X' and here $F(X|X^n)$ denotes the posterior cumulative distribution of $P_{X|X^n}$. This quantity is equal to

$$\begin{aligned} E[I_{T_1} - I_{T_2}] = \omega_0^2 \iint (I_{T_1}(X, X') - I_{T_2}(X, X'))dF(X)dF(X') \\ + 2\sum_{i=1}^{n} \omega_0\omega_i \int (I_{T_1}(X_i, X') - I_{T_2}(X_i, X'))dF(X') \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{T_1}(X_i, X_j) - I_{T_2}(X_i, X_j)), \end{aligned}$$

where we have exploited the fact that X' has the same distribution as X and the same observations. Given $(\omega_0, \ldots, \omega_n)$, it can be seen that the first two terms depend on the prior distribution $F \sim Dp(s, \alpha^*)$ and the last term is only due to the observations.

Theorem 3. *The variables Y and Z are said to be concordant (dependent) with posterior probability $(1-\gamma)$ provided that*

$$\mathcal{P}(E[I_{T_1} - I_{T_2}]/2 > 0 | X^n) > (1-\gamma)\,, \tag{27}$$

and they are said to be discordant provided that

$$\mathcal{P}(E[I_{T_1} - I_{T_2}]/2 < 0 | X^n) > (1-\gamma), \tag{28}$$

where $\mathcal{P}$ is the probability computed with respect to $(\omega_0, \omega_1, \ldots, \omega_n) \sim Dir(s, 1, \ldots, 1)$ and $dF \sim Dp(s, \alpha^*)$. Finally, they are said to be simply dependent with posterior probability $(1-\gamma)$ provided that

$$0 \notin (1-\gamma) HDI(E[I_{T_1} - I_{T_2}]/2 | X^n), \tag{29}$$

where HDI denotes the posterior Highest Density Interval of $E[I_{T_1} - I_{T_2}]/2$.

The divisor 2 in $E[I_{T_1} - I_{T_2}]/2$ is only a scaling factor so that the expectation lies in $[-0.5, 0.5]$. The theorem simply follows from the fact that $E[I_{T_1} - I_{T_2}]$ is the same measure of dependence used in Kendall's τ test. In this respect, it is worth to highlight the connection with Kendall's τ. By exploiting the properties of DP, we have that the posterior mean of $E[I_{T_1} - I_{T_2}]$ for large n is approximately equal to.

$$\mathcal{E}(E[I_{T_1} - I_{T_2}] | X^n) \approx \frac{1}{(n+1)n} \sum_{i=1}^{n} \sum_{j=1}^{n} (I_{T_1}(X_i, X_j) - I_{T_2}(X_i, X_j)) \tag{30}$$

and this is exactly Kendall's sample τ coefficient. In fact, Kendall's sample τ coefficient is defined as:

$$T = 2 \sum_{1 \le i < j \le n} \frac{A_{ij}}{n(n-1)} = 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{A_{ij}}{n(n-1)}, \tag{31}$$

with

$$A_{ij} = \begin{cases} 1, & \text{if } (Y_i - Y_j)(Z_i - Z_j) > 0, \\ -1, & \text{if } (Y_i - Y_j)(Z_i - Z_j) < 0. \end{cases}$$

Observe that T can also be rewritten as:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{A_{ij}}{n(n-1)}, \tag{32}$$

in terms of all the A_{ij} pairs, which is proportional to (30) for large n. This clarifies the connection between our Bayesian test of dependence for continuous variables based on $E[I_{T_1} - I_{T_2}]/2$ and Kendall's τ test.

As for the dependence test for binary variables, we will make inferences using IDP. Inferences with IDP can computed by verifying if

$$\underline{\mathcal{P}}(E[I_{T_1} - I_{T_2}]/2 > 0 | X^n) > (1-\gamma), \quad \overline{\mathcal{P}}(E[I_{T_1} - I_{T_2}]/2 > 0 | X^n) > (1-\gamma). \tag{33}$$

Theorem 4. *The upper probability* $\overline{\mathcal{P}}(E[I_{T_1} - I_{T_2}]/2 > 0|X^n)$ *is obtained by a prior measure* $\alpha^* = 0.5\delta_{X_0^a} + 0.5\delta_{X_0^b}$ *with* $X_0^a > X_0^b > X_i$ *for* $i = 1, \ldots, n$. *The lower probability* $\underline{\mathcal{P}}(E[I_{T_1} - I_{T_2}]/2 > 0|X^n)$ *is obtained by a prior measure* $\alpha^* = 0.5\delta_{X_0^a} + 0.5\delta_{X_0^b}$ *with* $X_0 = (Y_0, Z_0)$ *and* $Y_0^a > Y_0^b > Y_i$ *and* $Z_0^a < Z_0^b < Z_i$ *for* $i = 1, \ldots, n$.

Proof. We have that

$$\begin{aligned} E[I_{T_1} - I_{T_2}] = \omega_0^2 \iint (I_{T_1}(X, X') - I_{T_2}(X, X'))dF(X)dF(X') \\ + 2\sum_{i=1}^{n} \omega_0\omega_i \int (I_{T_1}(X_i, X') - I_{T_2}(X_i, X'))dF(X') \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{T_1}(X_i, X_j) - I_{T_2}(X_i, X_j)). \end{aligned}$$

We want to maximize $I_{T_1}(X, X')$. Since $\iint I_{T_1}(X, X')\delta_{X_0^a}(X)\delta_{X_0^a}(X')dXdX' = 0$, we need at least two Dirac's deltas. Hence, we consider the mixture $dF = m\delta_{X_0^a} + (1 - m)\delta_{X_0^b}$ with $X_0^a > X_0^b > X_i$ for $i = 1, \ldots, n$. Then we have that

$$E[I_{T_1} - I_{T_2}] = m(1 - m)\omega_0^2 + 2\sum_{i=1}^{n} \omega_0\omega_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{T_1}(X_i, X_j) - I_{T_2}(X_i, X_j)),$$

and so we have maximized the second term. For the first term depending on $m(1 - m)$, the maximum is obtained at $m = 1/2$. For the lower probability, the proof is similar. □

The lower and upper HDI intervals can also be obtained as in Theorem 4. Again in this case, the value of s determines how quickly lower and upper posteriors converge as the number of observations increases. We choose $s = 0.5$ as for the binary test.

Example 2. *Also in this case we consider three matrices of* 10 *paired continuous i.i.d. observations*

$$\begin{aligned} X_a^{10} &= \begin{bmatrix} -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & -0.3 & -0.2 & -0.1 \end{bmatrix}^T, \\ X_b^{10} &= \begin{bmatrix} -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & -0.4 & -0.3 & -0.2 & -0.1 \end{bmatrix}^T, \\ X_c^{10} &= \begin{bmatrix} -0.1 & 0.2 & 0.3 & -0.4 & -0.5 & -0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & -0.2 & -0.3 & -0.4 & 0.5 & 0.5 & 0.4 & -0.3 & -0.2 & -0.1 \end{bmatrix}^T. \end{aligned} \tag{34}$$

They correspond to different degrees of dependence. Figure 2 shows the lower and upper posteriors for the three cases a, b, c *and the relative* HDI *intervals at* 95% *probability (the filled in areas). In case (a), the two variables are dependent (concordant) with probability greater than* 0.95, *since all the mass of the lower and upper distributions are in the interval* $[0, 0.5]$. *In the second case, we are in an indeterminate situation, that is, the lower and upper are in disagreement, which means that the inference is prior dependent. In the third case, we can only say that they are not dependent at* 95% *since both the* HDI *intervals include the zero.*

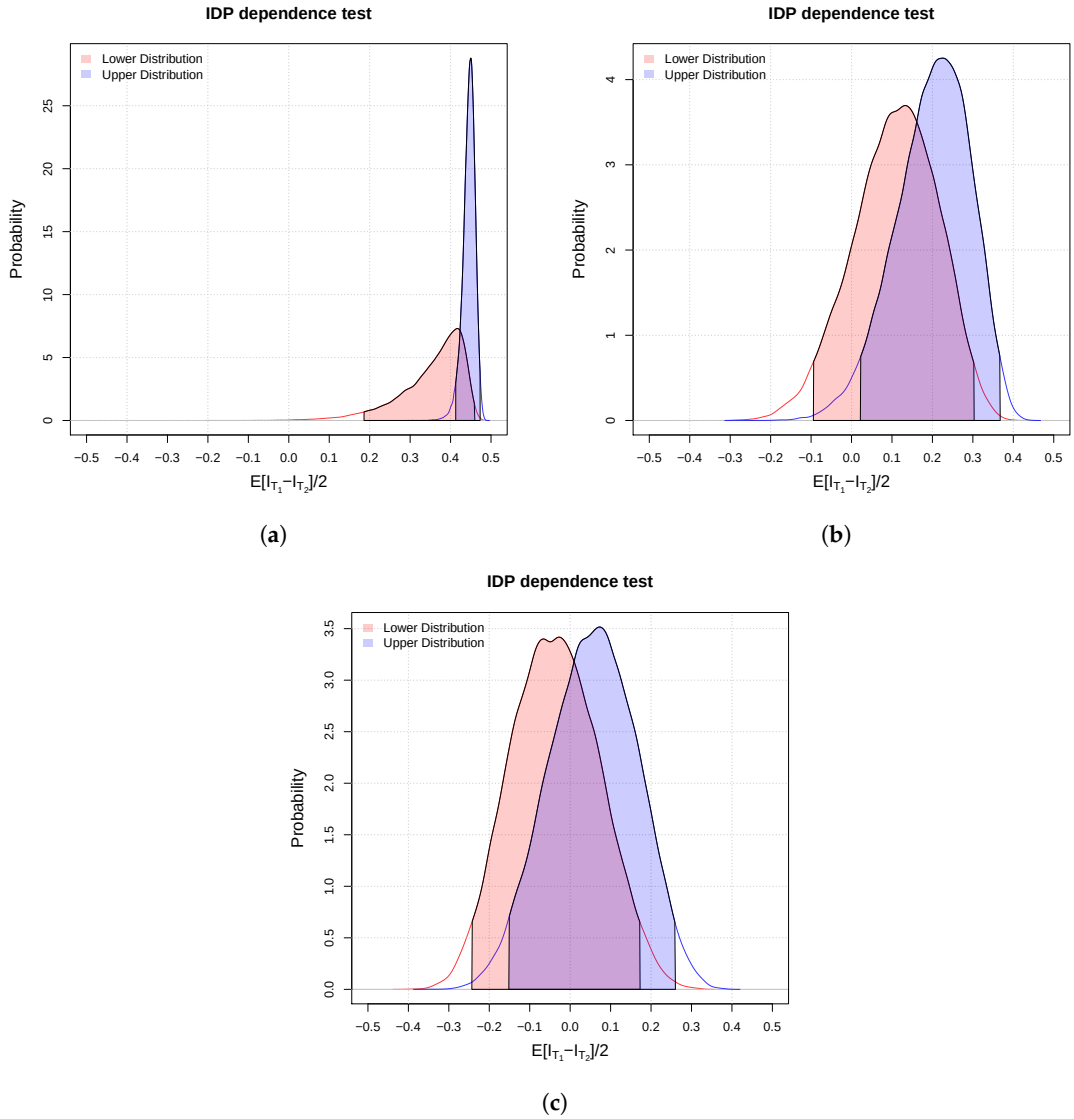


Figure 2. Three possible results of the independence hypothesis testing for continuous variables. The red and blue filled areas correspond respectively to the lower and upper HDI. (**a**) Dependent at 95%; (**b**) Indeterminate at 95%; (**c**) They are not dependent at 95%.

3.3. Bayesian Bivariate Independence Test for Mixed Continuous-Binary Variables

Let us assume that the variables $Y \in \mathbb{R}$ and $Z \in \{0,1\}$. Our aim is to devise a Bayesian independence test based on the DP. We introduce the auxiliary variable X' as done before. To derive our test, we start by defining the following indicator:

$$I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z').$$

This indicator is one if $X = (Y,0)$ and $X' = (Y',1)$, with $Y > Y'$ and zero otherwise. We can compute

$$E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] = \iint (I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z'))dF(X|X^n)dF(X'|X'^n), \quad (35)$$

where we have exploited the independence of X, X'. $F(X|X^n)$ denotes the posterior cumulative distribution of $P_{X|X^n}$. This quantity is equal to

$$\begin{aligned} E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] = \omega_0^2 \iint (I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z'))dF(X)dF(X') \\ + \sum_{i=1}^{n} \omega_0\omega_i \int (I_{(Y',\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z'))dF(X') \\ + \sum_{i=1}^{n} \omega_0\omega_i \int (I_{(Y_i,\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z_i))dF(X) \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{(Y_j,\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z_j)). \end{aligned}$$

For large n, we have that

$$\mathcal{E}(E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')]|X^n) \approx \frac{1}{(n+1)n}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{(Y_j,\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z_j), \tag{36}$$

which is equal to the rank of Y in the observations $(Y,0)$ with respect to the observations $(Y,1)$. Therefore, our dependence test is rank-based. It is clear that, in the case of independence of the variables Y and Z, the mean rank is equal to 0.125. Hence, we can formulate an independence test for mixed variables.

Theorem 5. *The variables Y and Z are dependent with posterior probability $(1-\gamma)$ provided that*

$$0 \notin (1-\gamma)HDI(4E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] - 0.5|X^n), \tag{37}$$

where HDI denotes the posterior Highest Density Interval of $4E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] - 0.5$.

The theorem follows from the fact that in case of independence between variables Y and Z the mean rank (36) scaled by 4 and shifted of -0.5 is equal to 0. Also in this case, we make inferences using IDP.

Theorem 6. *The upper probability $\overline{\mathcal{P}}(4E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] - 0.5 > 0|X^n)$ is obtained by a prior measure $\alpha^* = m\delta_{X_0^a} + (1-m)\delta_{X_0^b}$ with X_0^a equal to $(-\infty, 1)$ and X_0^b equal to $(\infty, 0)$ and*

$$m = \begin{cases} 0 & \text{if } \omega_0 + e_0 < e_1\,, \\ 1 & \text{if } e_1 < e_0 - \omega_0\,, \\ \dfrac{\omega_0 + e_0 - e_1}{2\omega_0} & \text{other,} \end{cases}$$

with $e_0 = \sum_{i=1}^{n} \omega_i I_{\{0\}}(Z_i)$ and $e_1 = \sum_{i=1}^{n} \omega_i I_{\{1\}}(Z_i)$. The lower probability $\underline{\mathcal{P}}(4E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] - 0.5 > 0|X^n)$ is obtained by a prior measure $\alpha^* = \delta_{X_0}$ with X_0 equal to $(-\infty, 0)$.

Proof. Consider the quantity

$$\begin{aligned} E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] = \omega_0^2 \iint (I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z'))dF(X)dF(X') \\ + \sum_{i=1}^{n} \omega_0\omega_i \int (I_{(Y',\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z'))dF(X') \\ + \sum_{i=1}^{n} \omega_0\omega_i \int (I_{(Y_i,\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z_i))dF(X) \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{(Y_j,\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z_j)), \end{aligned}$$

and $\alpha^* = m\delta_{X_0^a} + (1-m)\delta_{X_0^b}$ with X_0^a equal to $(-\infty, 1)$ and X_0^b equal to $(\infty, 0)$. Thus

$$\begin{aligned} E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] = m(1-m)\omega_0^2 + m\sum_{i=1}^{n} \omega_0\omega_i I_{\{0\}}(Z_i) + (1-m)\sum_{i=1}^{n} \omega_0\omega_i I_{\{1\}}(Z_i)) \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j (I_{(Y_j,\infty)}(Y_i)I_{\{0\}}(Z_i)I_{\{1\}}(Z_j). \end{aligned}$$

By computing the derivative

$$\frac{d}{dm} E[I_{(Y',\infty)}(Y)I_{\{0\}}(Z)I_{\{1\}}(Z')] = \omega_0 - 2\omega_0 m + e_0 - e_1 = 0,$$

we have that $m = \frac{\omega_0 + e_0 - e_1}{2\omega_0}$. The result is obtained by exploiting the fact that $m \in [0,1]$. For the lower probability, the computation is straightforward. □

The lower and upper HDI intervals can also be obtained as in Theorem 4. We choose $s = 0.5$ as for the previous tests.

Example 3. *We consider three matrices of* 10 *paired binary-continuous i.i.d. observations*

$$\begin{aligned} X_a^{10} &= \begin{bmatrix} -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \\ X_b^{10} &= \begin{bmatrix} -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T, \\ X_c^{10} &= \begin{bmatrix} -0.1 & -0.2 & -0.3 & -0.4 & -0.5 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}^T. \end{aligned} \tag{38}$$

Again, they correspond to different degrees of dependence. Figure 3 shows the lower and upper posteriors for the three cases a, b, c and the relative HDI intervals at 95% *probability (the filled in areas). In case (a), the two variables are dependent (concordant) with probability greater than* 0.95*, since all the mass of the lower and upper distributions are in the interval* $[0, 0.5]$*. In the second case, we are in an indeterminate situation, that is, the lower and upper are in disagreement. In the third case, we can only say that they are not dependent at* 95% *since both the HDI intervals include the zero.*

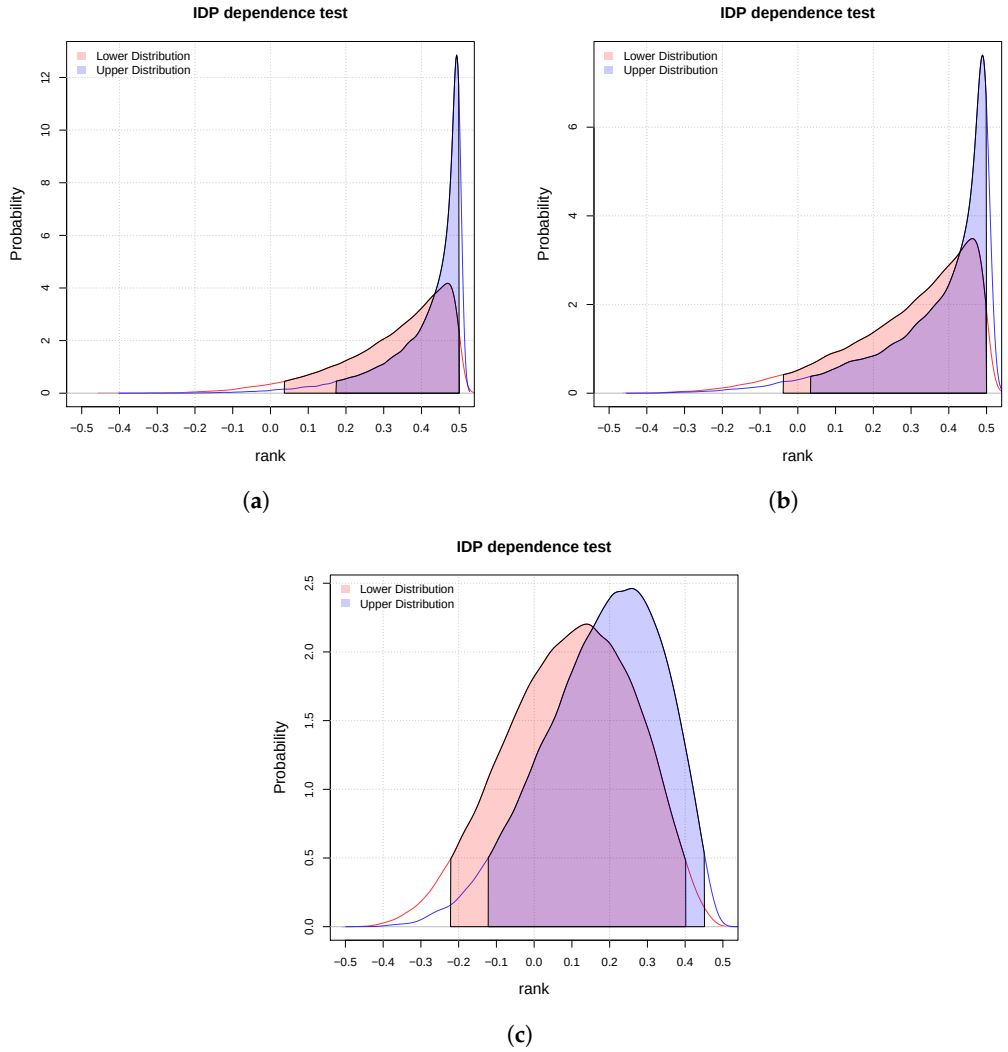


Figure 3. Three possible results of the independence hypothesis testing for pairs binary-continuous. The red and blue filled areas correspond respectively to the lower and upper HDI. (**a**) Dependent at 95%; (**b**) Indeterminate at 95%; (**c**) They are not dependent at 95%.

4. Experiments

We compare our Bayesian testing approach in the three discussed main scenarios where both variables are binary, both are continuous and one is binary and the other is continuous. The goal is to decide whether the two variables are dependent or independent. We generate n samples ($n = 20$ and 50) using the distributions defined in Table 1. Ten thousand repetitions are used by forcing the variables to be independent (so $\beta = 0$) and thousand repetitions where the variables are dependent, for each value of $\beta > 0$. The value of β is varied as explained in the table. For each n, β and each of these twenty thousand samples (for which we know the correct result of the test), we run the new approach versus χ^2 test, Kendall τ test and Kolmogorov–Smirnov test, respectively for the binary-binary, continuous-continuous and binary-continuous cases. For each run of each method, we record their p-values, while for the new approach we compute γ corresponding to the limiting credible region $1 - \gamma$ wide where the decision changes between dependent and independent. Such value is related to the p-values of the other tests and can be used for decision making by comparing it against a threshold (just as it is done with the p-values). However, it should be observed that thresholds different from 0.05 or 0.01 are hardly used in practice in null hypothesis significance tests. Conversely, for a Bayesian tests $1 - \gamma$ is a probability and, therefore, we can

take decisions with probability 0.99, 0.95 but also 0.7 or even 0.51 depending on the application (and the loss function). However, instead of fixing a threshold (which is a subjective choice) to decide between the options dependent and non-dependent with probability $1-\gamma$, we use Receiver Operating Characteristic (ROC) curves. ROC curves give the quality of the approaches for all possible thresholds. The curves are calculated as usual by varying the threshold from 0 to 1 and computing the sensitivity (or true positive rate) and specificity (or one minus false positive rate) (this is slightly different from the common approach of drawing ROC curves as a function of the true positive rate and false positive rate [26–28]). ROC curves are always computed considering different degrees of dependence (different values for $\beta \neq 0$) against independence ($\beta = 0$). We apply the same criterion to p-values for comparing the methods across a wider range of decision criteria. We have used the R package "pROC" to compute the ROC curves [29].

Table 1. Data generation setup. In order to generate independent data, β is set to zero. Larger values of β increase their dependency.

Variable 1	Variable 2	Distribution
Binary	Binary	Multinomial distr. with $[P(00), P(01), P(10), P(11)] \propto [3, 3+\beta, 3+\beta, 3]$.
Continuous	Continuous	Bivariate Gaussian with means 0 and covariance matrix $\begin{bmatrix} 10 & \beta \\ \beta & 3 \end{bmatrix}$.
Binary	Continuous	Half of the samples have the binary variable set to zero and half to one. When that variable is zero, then for the continuous use $\Gamma(10, 2)$, otherwise $\Gamma(10+\beta, 2+\beta)$.

Figures 4–6 present the comparison of the new approach (which we name as *IBinary*, *ICont* or *IMixed* to explicitly account for the types of variables been analyzed) using $s \approx 0$ against the appropriate competitor. With such choice of s, the new approach runs without indeterminacy and can be directly compared against usual methods. As we see in the figures, the new method performs very similar to each competitor, with the advantage of being compatible among different types of data (the p-values of the other methods, among different data types, cannot be compared to each other). This is useful when one works with multivariate models involving multiple data types. As expected, the quality of the methods increases with the increase of β and of the sample size.

Figures 7–9 present the ROC curves for the methods χ^2, Kendall τ and Kolmogorov–Smirnov, respectively. These curves are separated according to whether the instance is considered determinate or indeterminate by the new approach. In other words, for each one of the twenty thousand repetitions, we run the corresponding usual test and then we check whether the output of the new approach is determinate or indeterminate (applying $s = 0.5$), and we split the instances accordingly (blue curves show the accuracy over instances that are considered easy (determinate cases) while green curves over instances that are hard (indeterminate cases)—we also present the overall accuracy of the method using red curves). As we see, such division is able to identify easy-to-classify and hard-to-classify cases, since the ROC curves for the cases deemed as indeterminate by the new approach suggest a performance not better than a random guess (green curves). using the new approach, This means that if we would devise another test (called "50/50 when indeterminate") which returns the same response as *IBinary*, *ICont* or *IMixed* when they are determinate, and issues a random answer (with 50/50 chance) otherwise, then this "50/50 when indeterminate" test would have the same ROC curve as χ^2, Kendall τ and Kolmogorov–Smirnov, respectively.

This suggests that the indeterminacy of IDP based tests is an additional useful information that our approach gives to the analyst. In these cases she/he knows that (i) her/his posterior decisions would depend on the choice of the prior DP measure; (ii) deciding between the two hypotheses under test is a difficult problem as shown by the comparison with the DP with $s = 0$, χ^2, Kendall τ and Kolmogorov–Smirnov. Based on this additional information, the analyst can for example decide to collect additional measurements to eliminate the indeterminacy (in fact when the number of observations goes to infinity the indeterminacy goes to zero).

This represents a second advantage of our IDP approach, once we have fixed the value of s (e.g., $s = 0.5$) it can automatically identify the risky cases where a decision must be taken with additional caution. For this reason, we suggest to use the IDP based test for dependence and not $s = 0$.

Finally, Tables 2–4 present the values for the *Area under the curve* (AUC) in Chaper 5 in [30] of the ROC curves discussed previously, as well as similar experimental setup but with different values of s: 0.25, 0.5 and 1. Table 2 has results for binary variable versus binary variable, Table 3 for continuous variable versus continuous variable, and Table 4 for continuous variable versus binary variable. Overall, results show that *IBinary* has similar performance as χ^2 test, *ICont* has similar performance as Kendall's τ test and *IMixed* is similar to Kolmogorov–Smirnov (KS) test. The most interesting outcome is the comparison, in each scenario, of the frequentist test over whole data, over only data samples that were considered determinate by the new test, and over only data samples that were considered indeterminate. We clearly see that the AUC values over the cases considered indeterminate are much inferior to the values over cases considered determinate, which indicates that the new test has a good ability to discriminate easy and hard cases. ROC curves for values of s other than 0.5 were omitted for clarity of exposition, but they are very similar to those obtained for $s = 0.5$.

Table 2. Area under the ROC curve (AUC) values for all the performed experiments using different values of s, β and n. *IBinary* shows the AUC for the new test applied to two binary variables and $s \approx 0$. The columns χ^2 test, Det.cases, and Indet.cases show the AUC obtained by the χ^2 test over all samples, only over samples considered determinate by *IBinary* (with the corresponding s) and finally only over samples considered indeterminate by *IBinary*.

s	n	β	*IBinary*	Chisq	Det.cases	Indet.cases
0.25	20	1	0.5562	0.5629	0.5653	0.4890
0.5	20	1	0.5544	0.5596	0.5645	0.5233
1	20	1	0.5491	0.5551	0.5642	0.5153
0.25	20	3	0.7341	0.7502	0.7567	0.4266
0.5	20	3	0.7388	0.7551	0.7686	0.4526
1	20	3	0.7330	0.7502	0.7717	0.4888
0.25	50	1	0.6372	0.6425	0.6449	0.5125
0.5	50	1	0.6319	0.6353	0.6393	0.4747
1	50	1	0.6366	0.6407	0.6492	0.4954
0.25	50	3	0.9145	0.9110	0.9127	0.5205
0.5	50	3	0.9130	0.9090	0.9115	0.4473
1	50	3	0.9134	0.9081	0.9123	0.5642

Table 3. Area under the ROC curve (AUC) values for all the performed experiments using different values of s, β and n. *ICont* shows the AUC for the new test applied to two continuous variables and $s \approx 0$. Kendall, Det.cases, and Indet.cases show the AUC obtained by Kendall's test over all samples, only over samples considered determinate by *ICont* (with the corresponding s) and finally only over samples considered indeterminate by *ICont*.

s	n	β	*ICont*	Kendall	Det.cases	Indet.cases
0.25	20	1	0.5826	0.5858	0.5898	0.5101
0.5	20	1	0.5708	0.5729	0.5804	0.4987
1	20	1	0.5744	0.5742	0.5914	0.5004
0.25	20	2	0.7524	0.7506	0.7558	0.5037
0.5	20	2	0.7535	0.7502	0.7574	0.5203
1	20	2	0.7488	0.7407	0.7596	0.5447
0.25	50	1	0.6825	0.6888	0.6917	0.5051
0.5	50	1	0.6782	0.6869	0.6935	0.5633
1	50	1	0.6871	0.6960	0.7087	0.5204
0.25	50	2	0.9343	0.9191	0.9197	0.4933
0.5	50	2	0.9339	0.9208	0.9207	0.5487
1	50	2	0.9361	0.9205	0.9192	0.5499

Table 4. Area under the ROC curve (AUC) values for all the performed experiments using different values of s, β and n. *IMixed* shows the AUC for the new test applied to one binary and one continuous variables and $s \approx 0$. Kolmogorov–Smirnov (KS), Det.cases, and Indet.cases show the AUC obtained by KS test over all samples, only over samples considered determinate by *IMixed* (with the corresponding s) and finally only over samples considered indeterminate by *IMixed*.

s	n	β	*IMixed*	KS	Det.cases	Indet.cases
0.25	20	1	0.6159	0.6118	0.6139	0.5386
0.5	20	1	0.6150	0.5943	0.5989	0.5594
1	20	1	0.6132	0.6004	0.6104	0.5532
0.25	20	2	0.7176	0.7358	0.7392	0.5254
0.5	20	2	0.7202	0.7091	0.7159	0.4937
1	20	2	0.7163	0.7091	0.7233	0.4928
0.25	50	1	0.6997	0.7091	0.7109	0.4447
0.5	50	1	0.6966	0.7106	0.7149	0.4213
1	50	1	0.7076	0.7135	0.7224	0.4455
0.25	50	2	0.8526	0.8816	0.8832	0.3278
0.5	50	2	0.8497	0.8790	0.8818	0.3044
1	50	2	0.8562	0.8923	0.8986	0.2934

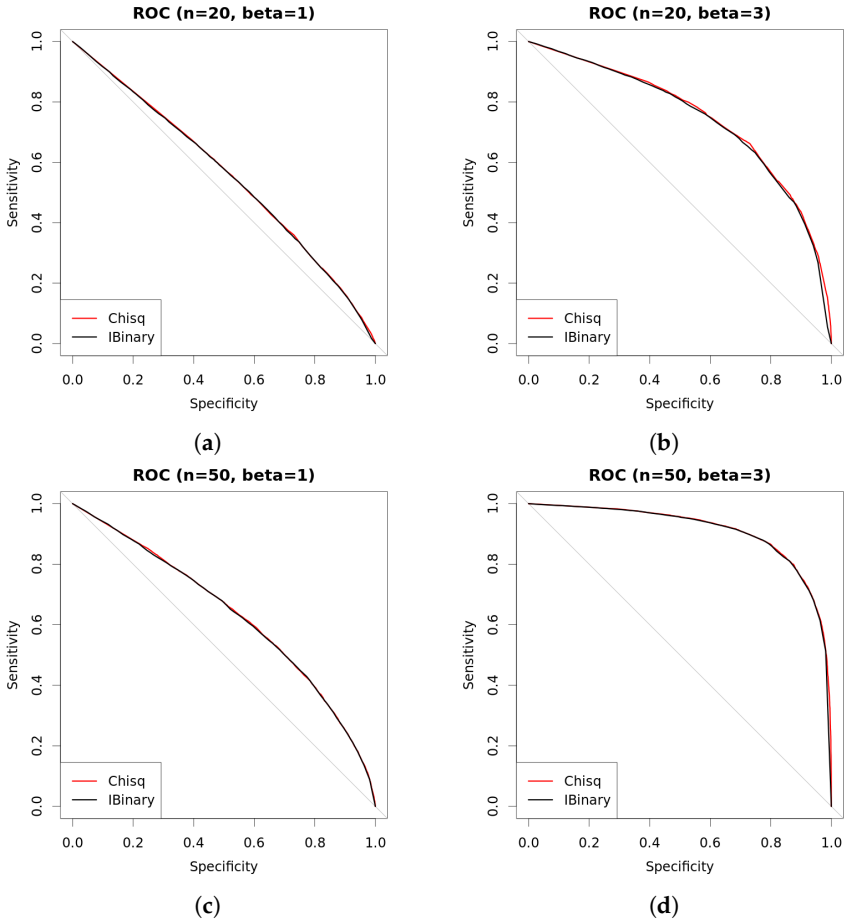


Figure 4. Comparison of approaches with binary data. New approach with $s \approx 0$ (so always determinate) is compared against χ^2 test using ROC curves. Curves are built using two thousand repetitions (one thousand where variables are independent ($\beta = 0$) and one thousand where they are dependent with β as shown in the figures). Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 3$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 3$).

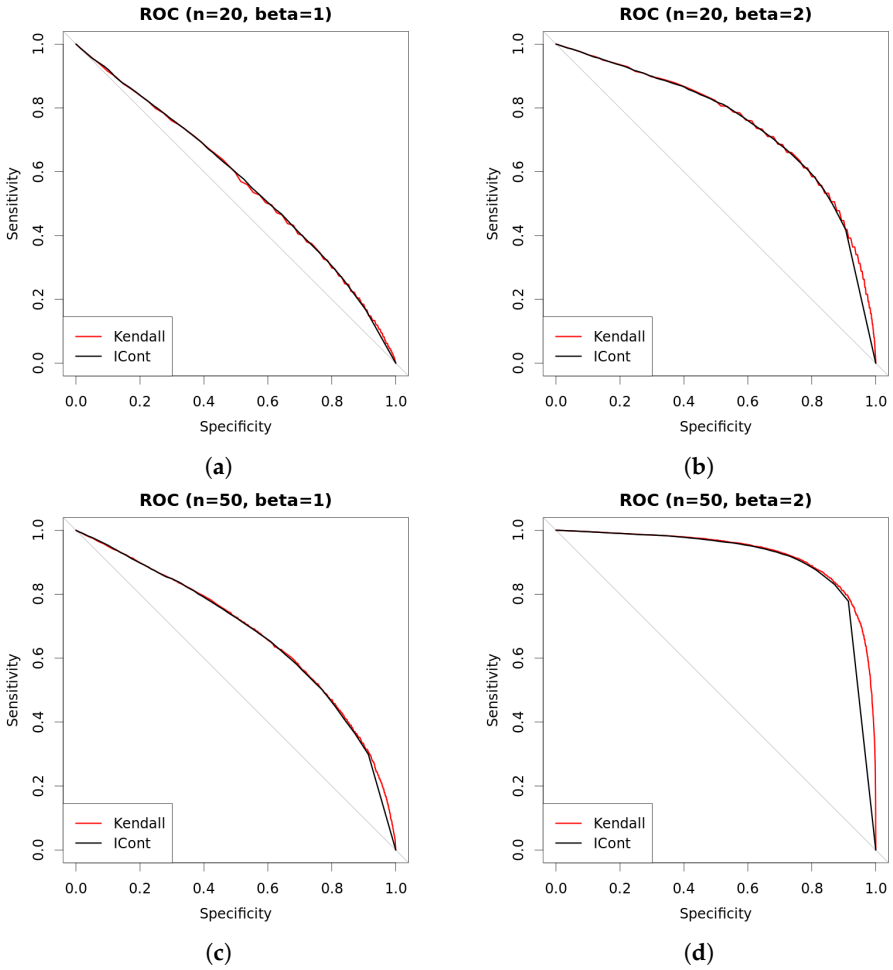


Figure 5. Comparison of approaches with continuous data. New approach with $s \approx 0$ (so always determinate) is compared against Kendall *tau* test using ROC curves. Curves are built using two thousand repetitions (one thousand where variables are independent ($\beta = 0$) and one thousand where they are dependent with β as shown in the figures). Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 2$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 2$).

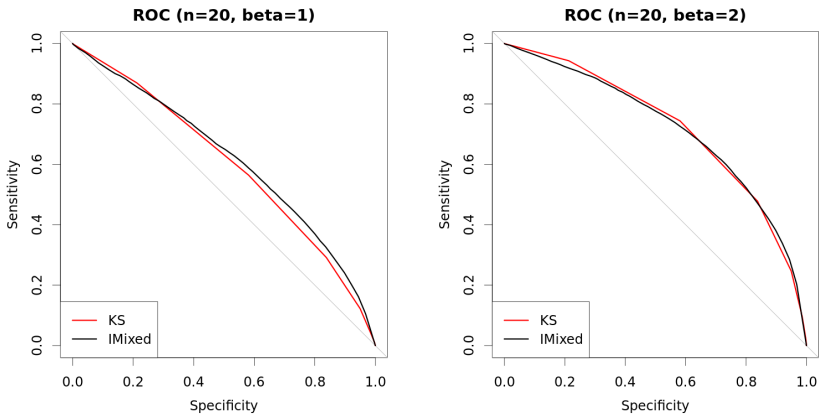


Figure 6. *Cont.*

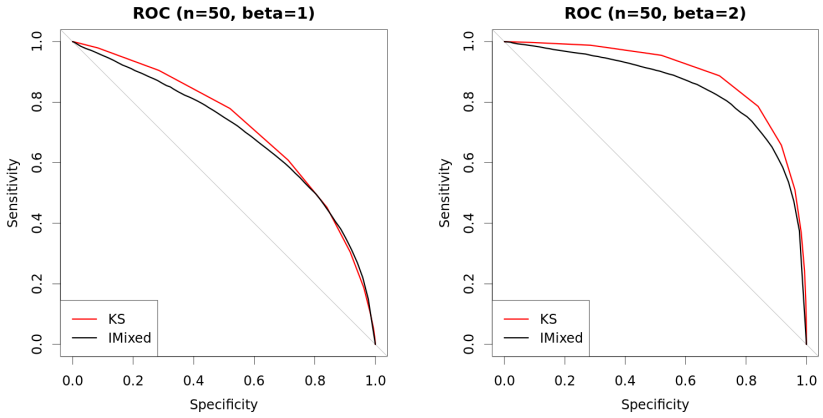


Figure 6. Comparison of approaches with mixed data. New method with $s \approx 0$ (so always determinate) is compared against Kolmogorov–Smirnov (KS) test using ROC curves. Curves are built using two thousand repetitions (one thousand where variables are independent ($\beta = 0$) and one thousand where they are dependent with β as shown in the figures). Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 2$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 2$).

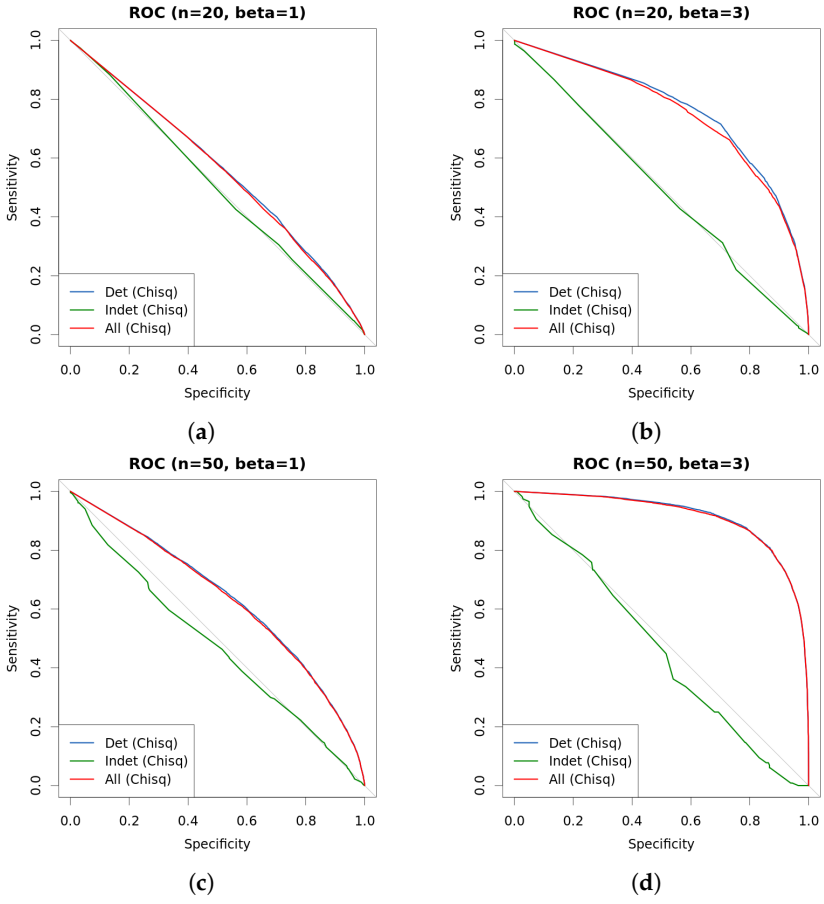


Figure 7. Comparison of approaches with binary data. New approach is used to differentiate instance by instance into hard-to-classify and easy-to-classify, and curves represent the outcome of χ^2 test under each such different scenarios. Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 3$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 3$).

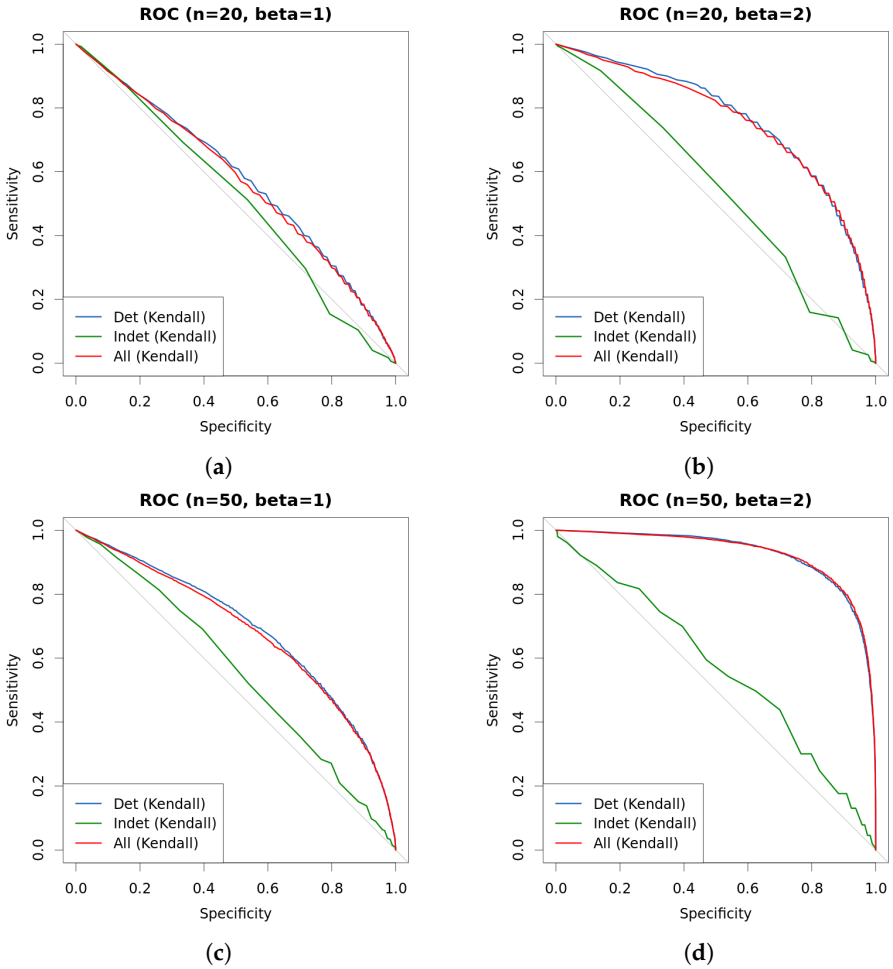


Figure 8. Comparison of approaches with continuous data. New approach is used to differentiate instance by instance into hard-to-classify and easy-to-classify, and curves represent the outcome of Kendall τ test under each such different scenarios. Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 2$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 2$).

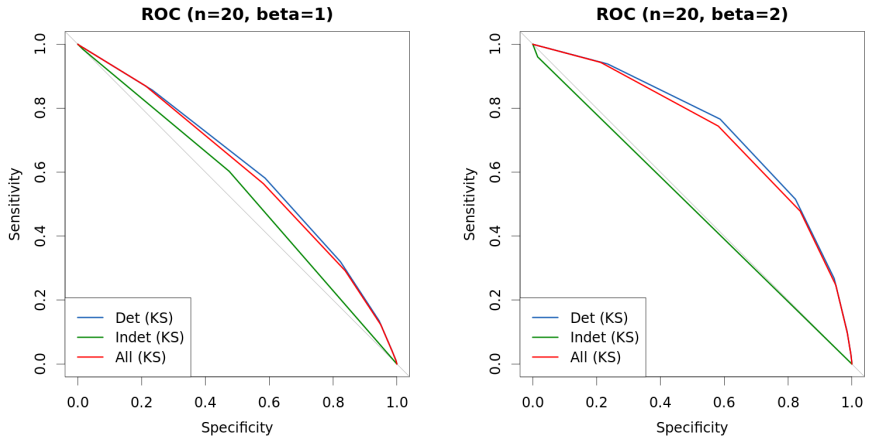


Figure 9. *Cont.*

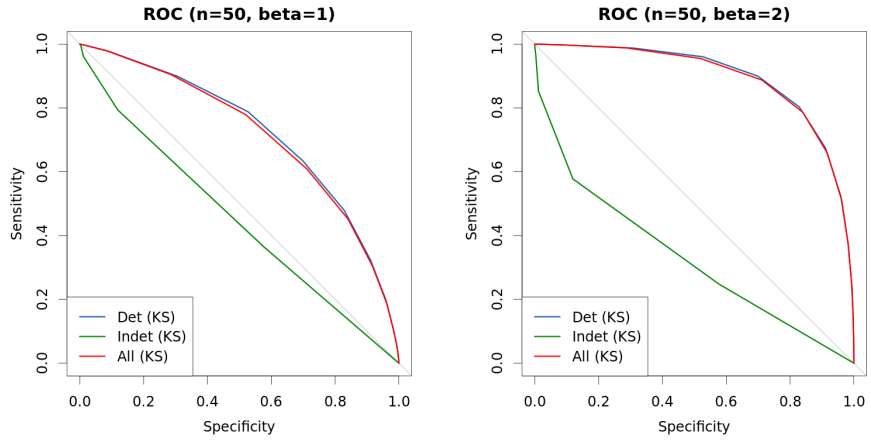


Figure 9. Comparison of approaches with mixed data. New approach is used to differentiate instance by instance into hard-to-classify and easy-to-classify, and curves represent the outcome of Kolmogorov–Smirnov (KS) test under each such different scenarios. Data are generated as explained in Table 1. (**a**) ROC ($n = 20$, $\beta = 1$); (**b**) ROC ($n = 20$, $\beta = 2$); (**c**) ROC ($n = 50$, $\beta = 1$); (**d**) ROC ($n = 50$, $\beta = 2$).

5. Conclusions

We have proposed three novel Bayesian methods for performing independence tests for binary, continuous and mixed binary-continuous variables. All of these tests are nonparametric and based on the Dirichlet Process. This has allowed us to use the same prior model for all the tests we have developed. Therefore, all the tests are "consistent", in the sense that the probabilities of dependence we compute with these tests are *commensurable* across the tests.

We have presented two versions of these tests: one based on a noninformative prior and one based on a conservative model of prior ignorance (IDP). Experimental results show that the prior ignorance method is more reliable than both the frequentist test and the noninformative Bayesian one, being able to isolate instances in which these tests are almost guessing at random. For future work, we plan to extend this approach in two directions: (1) feature selection in classification; (2) learning the structure (graph) of Bayesian networks and Markov Random Fields. The idea is to use our dependence tests to replace the frequentist tests that are commonly used for that purpose and evaluate the gain in terms of performance. For instance in case (1), we then could compare the accuracy of a classifier whose features are selected using our tests with that of a classifier whose features are selected by using frequentist tests. Our new approach is suitable since it addresses two limitations of currently used tests: they are based on null-hypothesis significance tests, and they cannot be applied to categorical and continuous variables at the same time in a commensurable way.

Author Contributions: All authors made substantial contributions to conception and design, data analysis and interpretation of data; all authors participate in drafting the article or revising it critically for important intellectual content; all authors gave final approval of the version to be submitted.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DP: Dirichlet Process
IDP: Imprecise Dirichlet Process

References

1. Raftery, A.E. Bayesian model selection in social research. *Sociol. Methodol.* **1995**, *25*, 111–164.
2. Goodman, S.N. Toward evidence-based medical statistics. 1: The P–value fallacy. *Ann. Intern. Med.* **1999**, *130*, 995–1004.
3. Kruschke, J.K. Bayesian data analysis. *Wiley Interdiscip. Rev. Cognit. Sci.* **2010**, *1*, 658–676.
4. Benavoli, A.; Mangili, F.; Ruggeri, F.; Zaffalon, M. Imprecise Dirichlet Process With Application to the Hypothesis Test on the Probability that $X \leq Y$. *J. Stat. Theory Pract.* **2015**, *9*, 658–684.
5. Benavoli, A.; Mangili, F.; Corani, G.; Zaffalon, M.; Ruggeri, F. A Bayesian Wilcoxon Signed-Rank Test Based on the Dirichlet Process. In Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, 21–26 July 2014; pp. 1026–1034.
6. Benavoli, A.; Corani, G.; Mangili, F.; Zaffalon, M. A Bayesian Nonparametric Procedure for Comparing Algorithms. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 1–9.
7. Mangili, F.; Benavoli, A.; de Campos, C.P.; Zaffalon, M. Reliable survival analysis based on the Dirichlet Process. *Biom. J.* **2015**, *57*, 1002–1019.
8. Kao, Y.; Reich, B.J.; Bondell, H.D. A nonparametric Bayesian test of dependence. **2015**, arXiv:1501.07198.
9. Nandram, B.; Choi, J.W. Bayesian analysis of a two-way categorical table incorporating intraclass correlation. *J. Stat. Comput. Simul.* **2006**, *76*, 233–249.
10. Nandram, B.; Choi, J.W. Alternative tests of independence in two-way categorical tables. *J. Data Sci.* **2007**, *5*, 217–237.
11. Nandram, B.; Bhatta, D.; Sedransk, J.; Bhadra, D. A Bayesian test of independence in a two-way contingency table using surrogate sampling. *J. Stat. Plan. Inference* **2013**, *143*, 1392–1408.
12. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163.
13. Blum, A.L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271.
14. Keogh, E.J.; Pazzani, M.J. Learning Augmented Bayesian Classifiers: A Comparison of Distribution-Based and Classification-Based Approaches. Available online: http://www.cs.rutgers.edu/∼pazzani/Publications/EamonnAIStats.pdf (accessed on 31 August 2016).
15. Jiang, L.; Cai, Z.; Wang, D.; Zhang, H. Improving Tree augmented Naive Bayes for class probability estimation. *Knowl. Based Syst.* **2012**, *26*, 239–245.
16. Ferguson, T.S. A Bayesian Analysis of Some Nonparametric Problems. *Ann. Stat.* **1973**, *1*, 209–230.
17. Ghosh, J.K.; Ramamoorthi, R. *Bayesian Nonparametrics*; Springer: Berlin/Heidelberg, Germany, 2003.
18. Rubin, D.B. Bayesian Bootstrap. *Ann. Stat.* **1981**, *9*, 130–134.
19. Walley, P. *Statistical Reasoning with Imprecise Probabilities*; Chapman & Hall: New York, NY, USA, 1991.
20. Coolen-Schrijner, P.; Coolen, F.P.; Troffaes, M.C.; Augustin, T. Imprecision in Statistical Theory and Practice. *J. Stat. Theory Pract.* **2009**, *3*, doi:10.1080/15598608.2009.10411907.
21. Augustin, T.; Coolen, F.P.; de Cooman, G.; Troffaes, M.C. *Introduction to Imprecise Probabilities*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
22. Berger, J.O.; Rios Insua, D.; Ruggeri, F. Bayesian Robustness. In *Robust Bayesian Analysis*; Insua, D.R., Ruggeri, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2000; Volume 152, pp. 1–32.
23. Berger, J.O.; Moreno, E.; Pericchi, L.R.; Bayarri, M.J.; Bernardo, J.M.; Cano, J.A.; De la Horra, J.; Martín, J.; Ríos-Insúa, D.; Betrò, B.; et al. An overview of robust Bayesian analysis. *Test* **1994**, *3*, 5–124.
24. Pericchi, L.R.; Walley, P. Robust Bayesian credible intervals and prior ignorance. *Int. Stat. Rev.* **1991**, *59*, doi:10.2307/1403571.
25. Dalal, S.; Phadia, E. Nonparametric Bayes inference for concordance in bivariate distributions. *Commun. Stat. Theory Methods* **1983**, *12*, 947–963.
26. Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Pearson Education: New York, NY, USA, 2006.
27. Jiang, L.; Li, C.; Cai, Z. Learning decision tree for ranking. *Knowl. Inf. Syst.* **2009**, *20*, 123–135.

28. Jiang, L.; Wang, D.; Zhang, H.; Cai, Z.; Huang, B. Using instance cloning to improve naive Bayes for ranking. *Int. J. Pattern Recognit. Artif. Intell.* **2008**, *22*, 1121–1140.
29. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, doi:10.1186/1471-2105-12-77.
30. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.