Entropy **2014**, *16*, 1376-1395; doi:10.3390/e16031376

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Bayesian Test of Significance for Conditional Independence: The Multinomial Model

Pablo de Morais Andrade *, Julio Michael Stern and Carlos Alberto de Bragança Pereira

Instituto de Matemática e Estatística, Universidade de São Paulo (IME-USP) Rua do Matão, 1010, Cidade Universitária, São Paulo, SP/Brasil, CEP: 05508-090; E-Mails: jstern@ime.usp.br (J.M.S.); cpereira@ime.usp.br (C.A.B.P.)

* Author to whom correspondence should be addressed; E-Mail: pablo.andrade@usp.br; Tel.:/Fax: +55-11-969783425.

Received: 3 December 2013; in revised form: 21 February 2014 / Accepted: 5 March 2014 / Published: 7 March 2014

Abstract: Conditional independence tests have received special attention lately in machine learning and computational intelligence related literature as an important indicator of the relationship among the variables used by their models. In the field of probabilistic graphical models, which includes Bayesian network models, conditional independence tests are especially important for the task of learning the probabilistic graphical model structure from data. In this paper, we propose the full Bayesian significance test for tests of conditional independence for discrete datasets. The full Bayesian significance test is a powerful Bayesian test for precise hypothesis, as an alternative to the frequentist's significance tests (characterized by the calculation of the *p*-value).

Keywords: hypothesis testing; probabilistic graphical models

---

1. Introduction

Barlow and Pereira [1] discussed a graphical approach to conditional independence. A probabilistic influence diagram is a directed acyclic graph (DAG) that helps model statistical problems. The graph is composed of a set of nodes or vertices, which represent the variables, and a set of arcs joining the nodes, which represent the dependence relationships shared by these variables.

The construction of this model helps us understand the problem and gives a good representation of the interdependence of the implicated variables. The joint probability of these variables can be written as a product of their conditional distributions, based on their independence and conditional independence.

The interdependence of the variables [2] is sometimes unknown. In this case, the model structure must be learned from data. Algorithms, such as the IC-algorithm (inferred causation) described in Pearl and Verma [3], have been designed to uncover these structures from the data. This algorithm uses a series of conditional independence tests (CI tests) to remove and direct the arcs, connecting the variables in the model and returning a DAG that minimally (with the minimum number of parameters and without loss of information) represents the variables in the problem.

The problem of constructing the DAG structures based on the data motivates the proposal of new powerful statistical tests for the hypothesis of conditional independence, because the accuracy of the structures learned is directly affected by the errors committed by these tests. Recently proposed structure learning algorithms [4–6] indicate that the results of CI tests are the main source of errors.

In this paper, we propose the full Bayesian significance test (FBST) as a test of conditional independence for discrete datasets. FBST is a powerful Bayesian test for a precise hypothesis and can be used to learn the DAG structures based on the data as an alternative to the CI tests currently in use, such as Pearson's chi-squared test.

This paper is organized as follows. In Section 2, we review the FBST. In Section 3, we review the FBST for the composite hypothesis. Section 4 gives an example of testing for conditional independence that can be used to construct a simple model with three variables.

2. The Full Bayesian Significance Test

The full Bayesian significance test was presented by Pereira and Stern [7] as a coherent Bayesian significance test for sharp hypotheses. In the FBST, the evidence for a precise hypothesis is computed.

This evidence is given by the complement of the probability of a credible set, called the *tangent* set, which is a subset of the parameter space in which the posterior density of each of the elements is greater than the maximum of the posterior density over the null hypothesis. This evidence is called the *e-value*, ev(H), and has many desirable properties as a statistical support. For example, Borges and Stern [8] described the following properties:

(1) provides a measure of significance for the hypothesis as a probability defined directly in the original parameter space.

(2) provides a smooth measure of the significance, both continuous and differentiable, of the hypothesis parameters.

(3) has an invariant geometric definition, independent of the particular parameterization of the null hypothesis being tested or the particular coordinate system chosen for the parameter space.

(4) obeys the likelihood principle.

(5) requires no *ad hoc* artifice, such as an arbitrary initial belief ratio between hypotheses.

(6) is a possibilistic support function, where the support of a logical disjunction is the maximum support among the support of the disjuncts.

(7) provides a consistent test for a given sharp hypothesis.

(8) provides compositionality operations in complex models.

(9) is an exact procedure, making no use of asymptotic approximations when computing the *e-value*.

(10) allows the incorporation of previous experience or expert opinions via prior distributions.

Furthermore, FBST is an exact test, whereas tests, such as the one presented in Geenens and Simar [9], are asymptotically correct. Therefore, the authors consider that a direct comparison between FBST and such test is not relevant in the context of this paper; considering, as future research, the comparison using small samples, in which case, FBST is still valid.

A more formal definition is given below.

Consider a model in a statistical space described by the triple, (Ξ, Δ, Θ), where Ξ is the sample space, Δ, the family of measurable subsets of Ξ and Θ the parameter space (Θ is a subset of $\Re^n$).

Define a subset of the parameter space, T_φ (tangent set), where the posterior density (denoted by f_x) of each element of this set is greater than φ.

$$T_\varphi = \{\theta \in \Theta | f_x(\theta) > \varphi\}. \tag{1}$$

The credibility of T_φ is given by its posterior probability,

$$\kappa = \int_{T_\varphi} f_x(\theta) d\theta = \int_\Theta f_x(\theta) \mathbb{1}_{T_\varphi}(\theta)\, d\theta, \tag{2}$$

where $\mathbb{1}_{T_\varphi}(\theta)$ is the indicator function.

$$\mathbb{1}_{T_\varphi}(\theta) = \begin{cases} 1 & \text{if } \theta \in T_\varphi \\ 0 & \text{otherwise} \end{cases}$$

Defining the maximum of the posterior density over the null hypothesis as f_x^*, with the maximum point at θ_0^*,

$$\theta_0^* \in \underset{\theta \in \Theta_0}{\operatorname{argmax}} f_x(\theta), \text{ and } f_x^* = f_x(\theta^*), \tag{3}$$

and defining $T^* = T_{f_x^*}$ as the tangent set to the null hypothesis, H_0, the credibility of T^* is κ^*.

The measure of the evidence for the null hypothesis (called the e-value), which is the complement of the probability of the set T^*, is defined as follows:

$$Ev(H_0) = 1 - \kappa^* = 1 - \int_\Theta f_x(\theta) \mathbb{1}_{T^*}(\theta)\, d\theta. \tag{4}$$

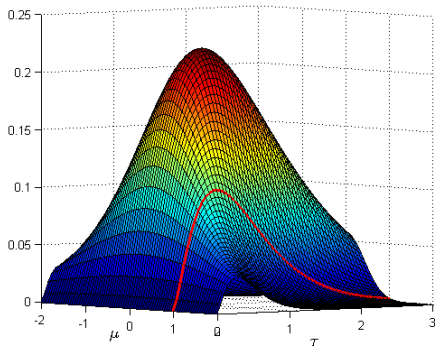
If the probability of the set, T^*, is large, the null set falls within a region of low probability, and the evidence is against the null hypothesis, H_0. However, if the probability of T^* is small, then the null set is in a region of high probability, and the evidence supports the null hypothesis.

2.1. FBST: Example of Tangent Set

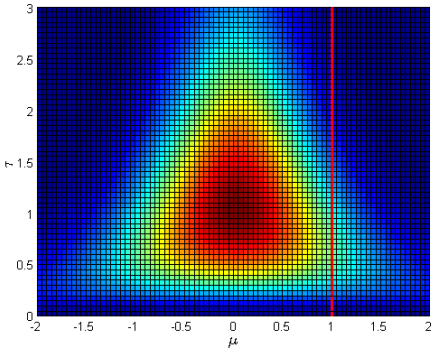
Figure 1 shows the tangent set for a null hypothesis $H_0 : \mu = 1$, for the posterior distribution, f_x, given bellow, where μ is the mean of a normal distribution and τ is the *precision* (the inverse of the variance $\tau = \frac{1}{\sigma^2}$):

$$f_x(\mu, \tau) \propto \tau^{1.5} e^{-\tau(\mu)^2 - 1.5\tau}. \tag{5}$$

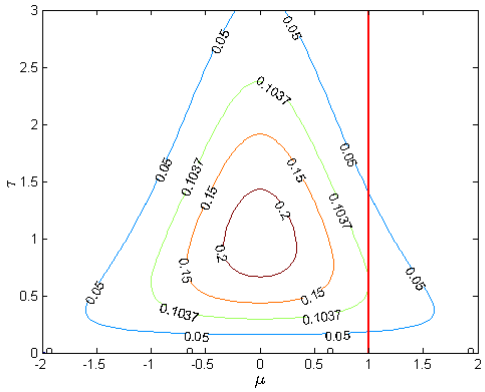
Figure 1. Example of a tangent set for the null hypothesis, $H_0 : \mu = 1.0$. In (**a**) and (**b**), the posterior distribution, f_x, is shown, with the red line representing the points in the null hypothesis ($\mu = 1$). In (**c**), the contours of f_x show that the points of maximum density in the null hypothesis, θ_0^*, have a density of 0.1037 ($f^* = f(\theta_0^*) = 0.1037$). The tangent set, T^*, of the null hypothesis, H_0, is the set of points inside the green contour line (points with a density greater than f^*), and the e-value of H_0 is the complement of the integral of f_x, as bounded by the green contour line.



(**a**) Posterior f_x. Red line: $\mu = 1.0$.



(**b**) Posterior f_x. Red line: $\mu = 1.0$.



(**c**) Contours of f_x. Red line: $\mu = 1.0$.

3. FBST: Compositionality

The relationship between the credibility of a complex hypothesis, H, and its elementary constituent, H_j, $j = 1, \ldots, k$, under the full Bayesian significance test, was analyzed by Borges and Stern [8].

For a given set of *independent* parameters, $(\theta_1, \ldots, \theta_k) \in (\Theta_1 \times \ldots \times \Theta_k)$, a complex hypothesis, H, can be given as follows:

$$H : \theta_1 \in \Theta_1^H \wedge \theta_2 \in \Theta_2^H \wedge \ldots \wedge \theta_k^H \in \Theta_k^H, \tag{6}$$

where Θ_j^H is a subset of the parameter space, Θ_j, for $j = 1, \ldots, k$ and is constrained to the hypothesis, H, which can be decomposed into its elementary components (hypotheses):

$$H_1 : \theta_1 \in \Theta_1^H$$
$$H_2 : \theta_2 \in \Theta_2^H$$
$$\cdots$$
$$H_k : \theta_k \in \Theta_k^H$$

The credibility of H can be evaluated based on the credibility of these components. The evidence in favor of the complex hypothesis, H (measured by its e-value), cannot be obtained directly from the evidence in favor of the elementary components; instead, it must be based on their *truth function*, W^j (or cumulative surprise distribution), as defined below. For a given elementary component (H_j) of the complex hypothesis, H, θ_j^* is the point of maximum density of the posterior distribution (f_x) that is constrained to the subset of the parameter space defined by hypothesis H_j:

$$\theta_j^* \in \underset{\theta_j \in \Theta_j^H}{\operatorname{argmax}} f_x(\theta_j) \text{ and } f_j^* = f_x\left(\theta_j^*\right). \tag{7}$$

The truth function, W_j, is the probability of the parameter subspace (region $R_j(v)$ of the parameter space defined below), where the posterior density is lower than or equal to the value, v:

$$R_j(v) = \{\theta_j \in \Theta_j | f_x(\theta_j) \leq v\},$$
$$W_j(v) = \int_{R_j(v)} f_x(\theta_j)\, d\theta_j. \tag{8}$$

The evidence supporting the hypothesis, H_j, is given as follows:

$$Ev(H_j) = W_j(f_j^*). \tag{9}$$

The evidence supporting the complex hypothesis can be then described in terms of the truth function of its components as follows.

Given two independent variables, X and Y, if $Z = XY$, with cumulative distribution functions $F_Z(z)$, $F_X(x)$ and $F_Y(y)$, then:

$$F_Z(z) = \Pr[Z \leq z] = \Pr[X \leq z/Y] = \int_0^\infty \Pr[X \leq z/y] f_Y(y)\, dy =$$
$$\int_0^\infty F_X(z/y) f_Y(y)\, dy = \int_0^\infty F_X(z/y) F_Y(dy). \tag{10}$$

Accordingly, we define a functional product for cumulative distribution functions, namely,

$$F_Z = F_X \otimes F_Y(z) = \int F_X(z/y) F_Y(dy). \tag{11}$$

The same result concerning the product of non-negative random variables can be expressed by the Mellin convolution of the probability density functions, as demonstrated by Kaplan and Lin [10], Springer [11] and Williamson [12].

$$f_Z(z) = (f_X \star f_Y)(z) = \int_0^\infty (1/y) f_X(z/y) f_Y(y)\, dy. \tag{12}$$

The evidence supporting the complex hypothesis can be then described as the Mellin convolution of the truth function of its components:

$$Ev(H) = W_1 \otimes W_2 \otimes W_3 \otimes \ldots \otimes W_k \left(f_1^* \cdot f_2^* \cdot f_3^* \cdot \ldots \cdot f_k^*\right). \tag{13}$$

The Mellin convolution of two truth functions, $W_1 \otimes W_2$, is the distribution function; see Borges and Stern [8]:

$$W_1 \otimes W_2(f_1^* \cdot f_2^*) = \int_0^\infty W_1\left(\frac{f_1^* \cdot f_2^*}{f}\right) W_2(df). \tag{14}$$

The Mellin convolution $W_1 \otimes W_2$ gives the distribution function of the product of two independent random variables, with distribution functions W_1 and W_2; see Kaplan and Lin [13] and Williamson [12]. Furthermore, the commutative and associative properties follow immediately for the Mellin convolution,

$$(W_1 \otimes W_2) \otimes W_3 = W_1 \otimes (W_2 \otimes W_3) = (W_1 \otimes W_3) \otimes W_2 = W_1 \otimes (W_3 \otimes W_2). \tag{15}$$

3.1. Mellin Convolution: Example

An example of a Mellin convolution to find the product of two random variables, Y_1 and Y_2, both of which have a Log-normal distribution, is given below.

Assume Y_1 and Y_2 to be continuous random variables, such that:

$$Y_1 \sim \ln\mathcal{N}\left(\mu_1, \sigma_1^2\right), \quad Y_2 \sim \ln\mathcal{N}\left(\mu_2, \sigma_2^2\right). \tag{16}$$

We denote the cumulative distributions of Y_1 and Y_2 by W_1 and W_2, respectively, i.e.,

$$W_1(y_1) = \int_{-\infty}^{y_1} f_{Y_1}(t)dt, \quad W_2(y_2) = \int_{-\infty}^{y_2} f_{Y_2}(t)dt, \tag{17}$$

where f_{Y_1} and f_{Y_2} are the density functions of Y_1 and Y_2, respectively. These distributions can be written as a function of two normally distributed random variables, X_1 and X_2:

$$\begin{aligned} \ln(Y_1) = X_1 &\sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), \\ \ln(Y_2) = X_2 &\sim \mathcal{N}\left(\mu_2, \sigma_2^2\right). \end{aligned} \tag{18}$$

We can confirm that the distribution of the product of these random variables $(Y_1 \cdot Y_2)$ is also Log-normal, using simple arithmetic operations:

$$\begin{aligned} Y_1 = e^{X_1} \text{ and } Y_2 &= e^{X_2}, \\ Y_1 \cdot Y_2 &= e^{X_1 + X_2}, \\ \ln(Y_1 \cdot Y_2) = X_1 + X_2 &\sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right), \\ \therefore Y_1 \cdot Y_2 &\sim \ln\mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right). \end{aligned} \tag{19}$$

The cumulative density function of $Y_1 \cdot Y_2$ $(W_{12}(y_{12}))$ is defined as follows:

$$W_{12}(y_{12}) = \int_{-\infty}^{y_{12}} f_{Y_1 \cdot Y_2}(t)dt, \tag{20}$$

where $f_{Y_1 \cdot Y_2}$ is the density function of $Y_1 \cdot Y_2$.

In the next section, we show different numerical methods for use in the convolution and condensation procedures, and we apply the results of these procedures to the example given here.

3.2. Numerical Methods for Convolution and Condensation

Williamson and Downs [14] developed the idea of probabilistic arithmetics. They investigated numerical procedures that allow for the computation of a distribution using arithmetic operations on random variables by replacing basic arithmetic operations on numbers with arithmetic operations on random variables. They demonstrated numerical methods for calculating the convolution of probability distributions for a set of random variables.

The convolution for the multiplication of two random variables, X_1 and X_2 ($Z = X_1 \cdot X_2$), can be written using their respective cumulative distribution functions, F_{X_1} and F_{Y_2}:

$$F_Z(z) = \int_0^z F_{X_1}\left(\frac{z}{t}\right) dF_{X_2}(t). \tag{21}$$

The algorithm for the numerical calculation of the distribution of the product of two independent random variables (Y_1 and Y_2), using their discretized marginal probability distributions (f_{Y_1} and f_{Y_2}) is shown in Algorithm 1 (an algorithm for a discretization procedure is given by Williamson and Downs [14]). The description of Algorithm 1 is given below.

(1) The algorithm has as inputs two discrete variables, Y_1 and Y_2, as well as their respective probabilistic density functions (pdf): f_{Y_1} and f_{Y_2}.

(2) The algorithm finds the products ($Y_1 \cdot Y_2$ and $f_{Y_1} \cdot f_{Y_2}$), resulting in N^2 bins, if f_{Y_1} and f_{Y_2} each have N bins.

(3) The values of $Y_1 \cdot Y_2$ are sorted in increasing order.

(4) The values of $f_{Y_1} \cdot f_{Y_2}$ are sorted according to the order of $Y_1 \cdot Y_2$.

(5) The cumulative density function (cdf) of the product $Y_1 \cdot Y_2$ is found (it has N^2 bins).

The numerical convolution of the two distributions with N bins, as described above, returns a distribution with N^2 bins. For a sequence of operations, such a large number of bins would be a problem, because the result of each operation would be larger than the input for the operations. Therefore, the authors have proposed a simple method for reducing the size of the output to N bins without introducing further error into the result. This operation is called *condensation* and returns the upper and lower bounds of each of the N bins for the distribution resulting from the convolution. The algorithm for the condensation process is shown in Algorithm 2. The description of Algorithm 2 is given below.

(1) The algorithm has as input a cdf with N^2 bins.

(2) For each group of N bins (there are N groups of N bins), the value of the cdf at the first bin is taken as the lower bound, and the value of the cdf at the last bin is taken as the upper bound.

(3) The algorithm returns a cdf with N bins, where each bin has a lower and an upper bound.

Algorithm 1 Find the distribution of the product of two random variables.

procedure CONVOLUTION($Y_1, Y_2, f_{Y_1}, f_{Y_2}$) ▷ Discrete pdf of Y_1 and Y_2
  $f \leftarrow array(0, size \leftarrow n^2)$ ▷ f and W has n^2 bins
  $W \leftarrow array(0, size \leftarrow n^2)$
  $y1y2 \leftarrow array(0, size \leftarrow n^2)$ ▷ keep $Y_1 * Y_2$
  for $i \leftarrow 1, n$ **do** ▷ f_1 and f_2 have n bins
    for $j \leftarrow 1, n$ **do**
      $f[(i-1) \cdot n + j] \leftarrow f_{Y_1}[i] \cdot f_{Y_2}[j]$
      $y1y2[(i-1) \cdot n + j] \leftarrow Y_1[i] * Y_2[j]$
    end for
  end for
  $sortedIdx \leftarrow order(y1y2)$ ▷ find order of $Y_1 * Y_2$
  $f \leftarrow f[sortedIdx]$ ▷ sort f according to $Y_1 * Y_2$
  $W[1] \leftarrow f[1]$
  for $i \leftarrow k, n^2$ **do** ▷ find cdf of $Y_1 \cdot Y_2$
    $W[k] \leftarrow f[k]$
    $W[k] \leftarrow W[k] + W[k-1]$
  end for
  return W ▷ Discrete cdf of $Y_1 \cdot Y_2$
end procedure

Algorithm 2 Find the upper lower bound for a cdf for condensation.

procedure HORIZONTALCONDENSATION(W) ▷ Histogram of a cdf with n^2 bins
  $W^l \leftarrow array(0, size \leftarrow n)$
  $W^u \leftarrow array(0, size \leftarrow n)$
  for $i \leftarrow 1, n$ **do**
    $W^l[i] \leftarrow W[(i-1) \cdot n + 1]$ ▷ lower bound after condensation
    $W^u[i] \leftarrow W[i \cdot n]$ ▷ upper bound after condensation
  end for
  return $\left[W^l, W^u\right]$ ▷ Histograms with upper/lower bounds
end procedure

3.2.1. Vertical Condensation

Kaplan and Lin [13] proposed a *vertical* condensation procedure for discrete probability calculations, where the condensation is done using the vertical axis, instead of the horizontal axis, as used by Williamson and Downs [14].

The advantage of this approach is that it provides greater control over the representation of the distribution; instead of selecting an interval of the domain of the cumulative distribution function (values assumed by the random variable) as a bin, we select the interval from the range of the cumulative distribution in $[0, 1]$, which should be represented by each bin.

In this case, it is also possible to focus on a specific region of the distribution. For example, if there is a greater interest in the behavior of the tail of the distribution, the size of the bins can be reduced in this region, consequently increasing the number of bins necessary to represent the tail of the distribution.

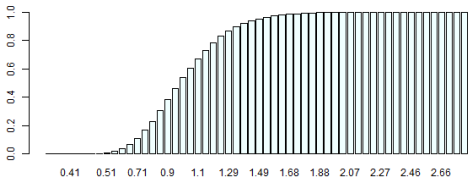
An example of such a convolution that is followed by a condensation procedure using both approaches is given in Section 3.1. For this example, we used discretization and condensation procedures, with the bins *uniformly* distributed over both axes. At the end of the condensation procedure, using the first approach, the bins are uniformly distributed *horizontally* (over the sample space of the variable). For the second approach, the bins of the cumulative probability distribution are uniformly distributed over the vertical axis on the interval $[0, 1]$. Algorithm 3 shows the condensation with the bins uniformly distributed over the vertical axis.

Algorithm 3 Condensation with the bins vertically uniformly distributed.

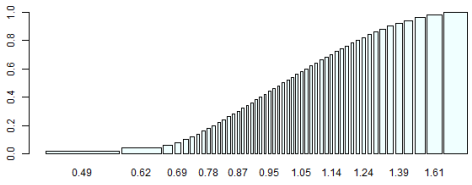
procedure VERTICALCONDENSATION(W, f, x) ▷ Histograms of a cdf and pdf, and breaks in the x-axis.
  $breaks \leftarrow [1/n, 2/n, ..., 1]$ ▷ uniform breaks in y-axis
  $W_n \leftarrow array\,(0, size \leftarrow n]$
  $x_n \leftarrow array\,(0, size \leftarrow n]$
  $lastbreak \leftarrow 1$
  $i \leftarrow 1$
  for all $b \in breaks$ **do**
    $w \leftarrow first(W \geq b)$ ▷ find break to create current bin
    if $W[w] \neq b$ **then** ▷ if the break is within a current bin
      $ratio \leftarrow (b - W[w-1])/(W[w] - W[w-1])$
      $x_n[i] \leftarrow \frac{1}{1/n}\,(sum\,(f[w-1] \cdot x[w-1]) + ratio \cdot f[w] \cdot x[w])$
      $W[i-1] \leftarrow b$
      $W_n[i] \leftarrow b$
      $f[i-1] \leftarrow f[w-1] + ratio \cdot f[w]$
      $f[i] \leftarrow (1 - ratio) \cdot f[w]$
    else
      $x_n[i] \leftarrow x[w]$
      $W_n[i] \leftarrow W[w]$
    end if
    $lastbreak \leftarrow b$
    $i \leftarrow i + 1$
  end for
  return $[W_n, x_n]$ ▷ Histograms with upper/lower bounds
end procedure

Figure 2 shows the cumulative distribution functions of Y_1 and Y_2 (Section 3.1) after they have been discretized with bins uniformly distributed over both the x- and y-axes (horizontal and vertical discretizations). Figure 3 shows an example of convolution followed by condensation (based on the example in Section 3.1), using both the horizontal and vertical condensation procedures and the true distribution of the product of two variables with Log-normal distributions.

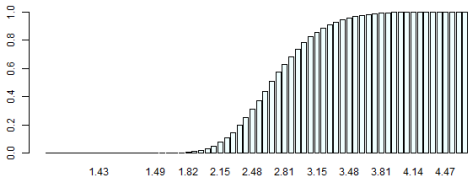
Figure 2. Example of different discretization methods for the representation of the cdf of two random variables (Y_1 and Y_2) with Log-normal distributions. In (**a**) and (**c**), respectively, the cdf of Y_1 and Y_2 are shown, with the bins uniformly distributed over the x-axis. In (**b**) and (**d**), respectively, the cdf of Y_1 and Y_2 are shown, with the bins uniformly distributed over the y-axis.



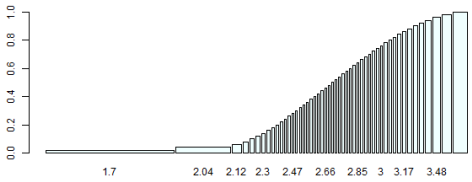
(**a**) W_1: Horizontal discretization



(**b**) W_1: Vertical discretization

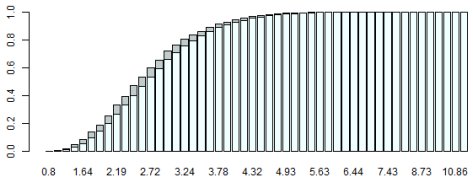


(**c**) W_2: Horizontal discretization

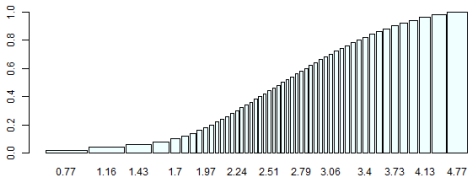


(**d**) W_2: Vertical discretization

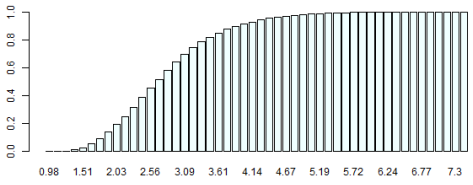
Figure 3. Example of the convolution of two random variables (Y_1 and Y_2) with Log-normal distributions. The result of the convolution $Y_1 \otimes Y_2$, followed by horizontal condensation (bins uniformly distributed over the x-axis), is shown in (**a**), and the result of vertical condensation (bins uniformly distributed over the y-axis) is shown in (**b**). The true distribution of the product $Y_1 \cdot Y_2$ is shown in (**c**) and (**d**), respectively, for the horizontal and vertical discretization procedures.



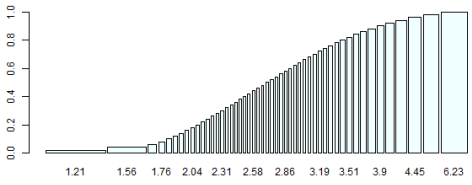
(**a**) $W_1 \otimes W_2$: Horizontal discretization



(**b**) $W_1 \otimes W_2$: Vertical discretization



(**c**) $Y_1 \cdot Y_2$: Horizontal discretization



(**d**) $Y_1 \cdot Y_2$: Vertical discretization

4. Test of Conditional Independence in Contingency Table Using FBST

We now apply the methods shown in the previous sections to find evidence of a complex null hypothesis of conditional independence for discrete variables.

Given the discrete random variables, X, Y and Z, with X taking values on $\{1, \ldots, k\}$ and Y and Z serving as categorical variables, the test for conditional independence $Y \perp\!\!\!\perp Z|X$ can be written as the complex null hypothesis, H:

$$H : [Y \perp\!\!\!\perp Z|X = 1] \wedge [Y \perp\!\!\!\perp Z|X = 2] \wedge \cdots \wedge [Y \perp\!\!\!\perp Z|X = k]. \tag{22}$$

The hypothesis, H, can be decomposed into its elementary components:

$$\begin{gathered} H_1 : Y \perp\!\!\!\perp Z|X = 1 \\ H_2 : Y \perp\!\!\!\perp Z|X = 2 \\ \cdots \\ H_k : Y \perp\!\!\!\perp Z|X = k \end{gathered}$$

Note that the hypotheses, $H_1, \ldots, H_k$, are *independent*. For each value, x, taken by X, the values taken by variables Y and Z are assumed to be random observations drawn from some distribution $p(Y, Z|X = x)$. Each of the elementary components is a hypothesis of independence in a contingency table. Table 1 shows the contingency table for Y and Z, which take values on $\{1, \ldots, r\}$ and $\{1, \ldots, c\}$, respectively.

Table 1. Contingency table of Y and Z for $X = x$ (hypothesis H_x); n_{yzx} is the count of $[Y, Z] = [y, z]$ when $X = x$.

	$Z = 1$	$Z = 2$	$\cdots$	$Z = c$
$Y = 1$	n_{11x}	n_{12x}	$\cdots$	n_{1cx}
$Y = 2$	n_{21x}	n_{22x}	$\cdots$	n_{2cx}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$Y = r$	n_{r1x}	n_{r2x}	$\cdots$	n_{rcx}

The test of the hypothesis, H_x, can be set up using the multinomial distribution for the cell counts of the contingency table and its natural conjugate prior, i.e., the Dirichlet distribution for the vector of the parameters $\theta_x = [\theta_{11x}, \theta_{12x}, \ldots, \theta_{rcx}]$.

For a given array of hyperparameters $\alpha_x = [\alpha_{11x}, \ldots, \alpha_{rcx}]$, the Dirichlet distribution is defined as:

$$f\left(\theta_x|\alpha_x\right) = \Gamma\left(\sum_{y,z}^{r,c} \alpha_{yzx}\right) \prod_{y,z}^{r,c} \frac{\theta_{yzx}^{\alpha_{yzx}-1}}{\Gamma\left(\alpha_{yzx}\right)}. \tag{23}$$

The multinomial likelihood for the given contingency table, assuming the array of observations $n_x = [n_{11x}, \ldots, n_{rcx}]$ and the sum of the observations $n_{..x} = \sum_{y,z}^{r,c} n_{yzx}$, is:

$$f\left(n_x|\theta_x\right) = n_{..x}! \prod_{y,z}^{r,c} \frac{\theta_{yzx}^{n_{yzx}}}{n_{yzx}!}. \tag{24}$$

The posterior distribution is thus a Dirichlet distribution, $f_n(\theta_x)$:

$$f_n\left(\theta_x\right) \propto \prod_{y,z}^{r,c} \theta_{yzx}^{\alpha_{yzx}+n_{yzx}-1}. \tag{25}$$

Under hypothesis H_x, we have $Y \perp\!\!\!\perp Z|X = x$. In this case, the joint distribution is equal to the product of the marginals: $p\left(Y = y, Z = z|X = x\right) = p\left(Y = y|X = x\right) p\left(Z = z|X = x\right)$. We can define this condition using the array of parameters, θ_x. In this case, we have:

$$H_x : \theta_{yzx} = \theta_{.zx} \cdot \theta_{y.x}, \forall y, z \tag{26}$$

where $\theta_{.zx} = \sum_y^r n_{yzx}$ and $\theta_{y.x} = \sum_z^c \theta_{yzx}$.

The elementary components of hypothesis H are as follows:

$$\begin{aligned} H_1 &: \theta_{yz1} = \theta_{.z1} \cdot \theta_{y.1}, \forall y, z \\ H_2 &: \theta_{yz2} = \theta_{.z2} \cdot \theta_{y.2}, \forall y, z \\ &\dots \\ H_k &: \theta_{yzk} = \theta_{.zk} \cdot \theta_{y.k}, \forall y, z \end{aligned} \tag{27}$$

The point of maximum density of the posterior distribution that is constrained to the subset of the parameter space defined by hypothesis H_x can be estimated using the maximum *a posteriori* (MAP) estimator under hypothesis H_x (the mode of parameters, θ_x). The maximum density (f_x^*) is the posterior density evaluated at this point.

$$\theta_{yzx}^* = \frac{n_{yzx}^{H_x} + \alpha_{yzx} - 1}{n_{..x}^{H_x} + \alpha_{..x} - r \cdot c} \text{ and } f_x^* = f_n(\theta_x^*), \tag{28}$$

where $\theta_x^* = [\theta_{11x}^*, \dots, \theta_{rcx}^*]$.

The evidence supporting H_x can be written in terms of the truth function, W_x, as defined in Section 3:

$$\begin{aligned} R_x(f) &= \{\theta_x \in \Theta_x | f_x\left(\theta_x\right) \leq f\}, \\ W_x(f) &= \int_{R_x(f)} f_n\left(\theta_x\right) d\theta_x \propto \int_{R_x(f)} \prod_{y,z}^{r,c} \theta_{yzx}^{\alpha_{yzx}+n_{yzx}-1} d\theta_x. \end{aligned} \tag{29}$$

The evidence supporting H_x is:

$$Ev(H_x) = W_x(f_x^*). \tag{30}$$

Finally, the evidence supporting the hypothesis of conditional independence (H) is given by the convolution of the truth functions that are evaluated at the product of the points of maximum posterior density, for each component of hypothesis H:

$$Ev(H) = W_1 \otimes W_2 \otimes \ldots \otimes W_k \left(f_1^* \cdot f_2^* \cdot \ldots \cdot f_k^*\right). \tag{31}$$

The e-value for hypothesis H can be found using modern mathematical integration methods. An example is given in the next section, using the numerical convolution, followed by the condensation procedures described in Section 3.2. Applying the horizontal condensation method results in an interval for the e-value (found using the lower and upper bounds resulting from the condensation process) and in a single value for the vertical procedure.

4.1. Example of CI Test Using FBST

In this section, we describe an example of the CI test using the full Bayesian significance test for conditional independence using samples from two different models. For both models, we test whether the variable, Y, is conditionally independent of Z given X.

Two probabilistic graphical models (M_1 and M_2) are shown in Figure 4, where the three variables, X, Y and Z, assume values in $\{1, 2, 3\}$. In the first model (Figure 4a), the hypothesis of independence $H : Y \perp\!\!\!\perp Z|X$ is true, but in the second model (Figure 4b), the same hypothesis is false. The synthetic conditional probability distribution tables (CPTs) used to generate the samples are given in Appendix.

Figure 4. Simple probabilistic graphical models. (**a**) Model M_1, where Y is conditionally independent of Z given X; (**b**) Model M_2, where Y is *not* conditionally independent of Z given X.

(**a**) $M_1 : Y \perp\!\!\!\perp Z|X$ (**b**) $M_2 : Y \not\perp\!\!\!\perp Z|X$

We calculate the intervals for the e-values and compare them, for hypothesis H of conditional independence, for both models: $Ev_{M_1}(H)$ and $Ev_{M_2}(H)$. The complexity hypothesis, H, can be decomposed into its elementary components:

$$H_1 : Y \perp\!\!\!\perp Z|X = 1$$
$$H_2 : Y \perp\!\!\!\perp Z|X = 2$$
$$H_3 : Y \perp\!\!\!\perp Z|X = 3$$

For each model, 5000 observations were generated; the contingency table of Y and Z for each value of X is shown in Table 2. The hyperparameters of the prior distribution were all set to one, because, in this case, the prior is equivalent to a uniform distribution (from Equation (23)):

$$\alpha_1 = \alpha_2 = \alpha_3 = [1, 1, 1],$$
$$f(\theta_1|\alpha_1) = f(\theta_3|\alpha_3) = f(\theta_3|\alpha_3) = 1. \tag{32}$$

The posterior distribution, found using Equations (24) and (25), is then given as follows:

$$f_n(\theta_1) \propto \prod_{y=1,z=1}^{3,3} \theta_{yz1}^{n_{yz1}}, f_n(\theta_2) \propto \prod_{y=1,z=1}^{3,3} \theta_{yz2}^{n_{yz2}}, f_n(\theta_3) \propto \prod_{y=1,z=1}^{3,3} \theta_{yz3}^{n_{yz3}}. \tag{33}$$

For example, for the given contingency table for Model M_1, when $X = 2$ (Table 2c), the posterior distribution is the following:

$$f_n(\theta_2) \propto \theta_{112}^{42} \cdot \theta_{122}^{41} \cdot \theta_{132}^{323} \cdot \theta_{212}^{39} \cdot \theta_{222}^{41} \cdot \theta_{232}^{341} \cdot \theta_{312}^{15} \cdot \theta_{322}^{21} \cdot \theta_{332}^{171}. \tag{34}$$

The point of highest density, in this example, following the hypothesis of independence (Equations (26) and (28)), was found to be the following:

$$\theta_2^* \approx [0.036, 0.039, 0.317, 0.038, 0.041, 0.329, 0.019, 0.020, 0.162]\,. \tag{35}$$

The truth function and the evidence supporting the hypothesis of independence given $X = 2$ (hypothesis H_2) for Model M_1, as given in Equations (29) and (30), are as follows:

$$\begin{aligned} R_2(f) &= \{\theta_2 \in \Theta_2 | f_n(\theta_2) \leq f\}\,, \\ W_2(f) &= \int_{R_2(f)} f_n(\theta_2)\, d\theta_2\,, \\ Ev_{M_1}(H_2) &= W_2(f_n(\theta_2^*)). \end{aligned} \tag{36}$$

We used the methods of numerical integration to find the e-value of the elementary components of hypothesis H (H_1,H_2 and H_3), and the results for each model are given below.

Table 2. Contingency tables of Y and Z for a given value of X for 5000 random samples. (**a**,**c**,**e**): samples from Model M_1 (Figure 4a) for $X = 1$, 2, and 3, respectively; (**b**,**d**,**f**): samples from Model M_2 (Figure 4b) for $X = 1$, 2, and 3, respectively.

(a) Model M_1 (for $X = 1$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	241	187	44	472
$Y=2$	139	130	30	299
$Y=3$	364	302	70	736
	744	619	144	1,507

(b) Model M_2 (for $X = 1$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	228	179	39	446
$Y=2$	25	33	211	269
$Y=3$	482	75	208	765
	735	287	458	1,048

(c) Model M_1 (for $X = 2$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	42	41	323	406
$Y=2$	39	41	341	421
$Y=3$	15	21	171	207
	96	103	835	1,034

(d) Model M_2 (for $X = 2$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	77	85	248	410
$Y=2$	165	135	120	420
$Y=3$	188	21	24	233
	430	241	392	1,036

(e) Model M_1 (for $X = 3$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	282	35	151	468
$Y=2$	131	37	79	247
$Y=3$	1,055	143	546	1,744
	1,468	215	776	2,459

(f) Model M_2 (for $X = 3$)

	$Z=1$	$Z=2$	$Z=3$	
$Y=1$	40	87	354	481
$Y=2$	119	104	27	250
$Y=3$	305	1,049	372	1,726
	464	1,240	753	2,457

E-values found using horizontal discretization:

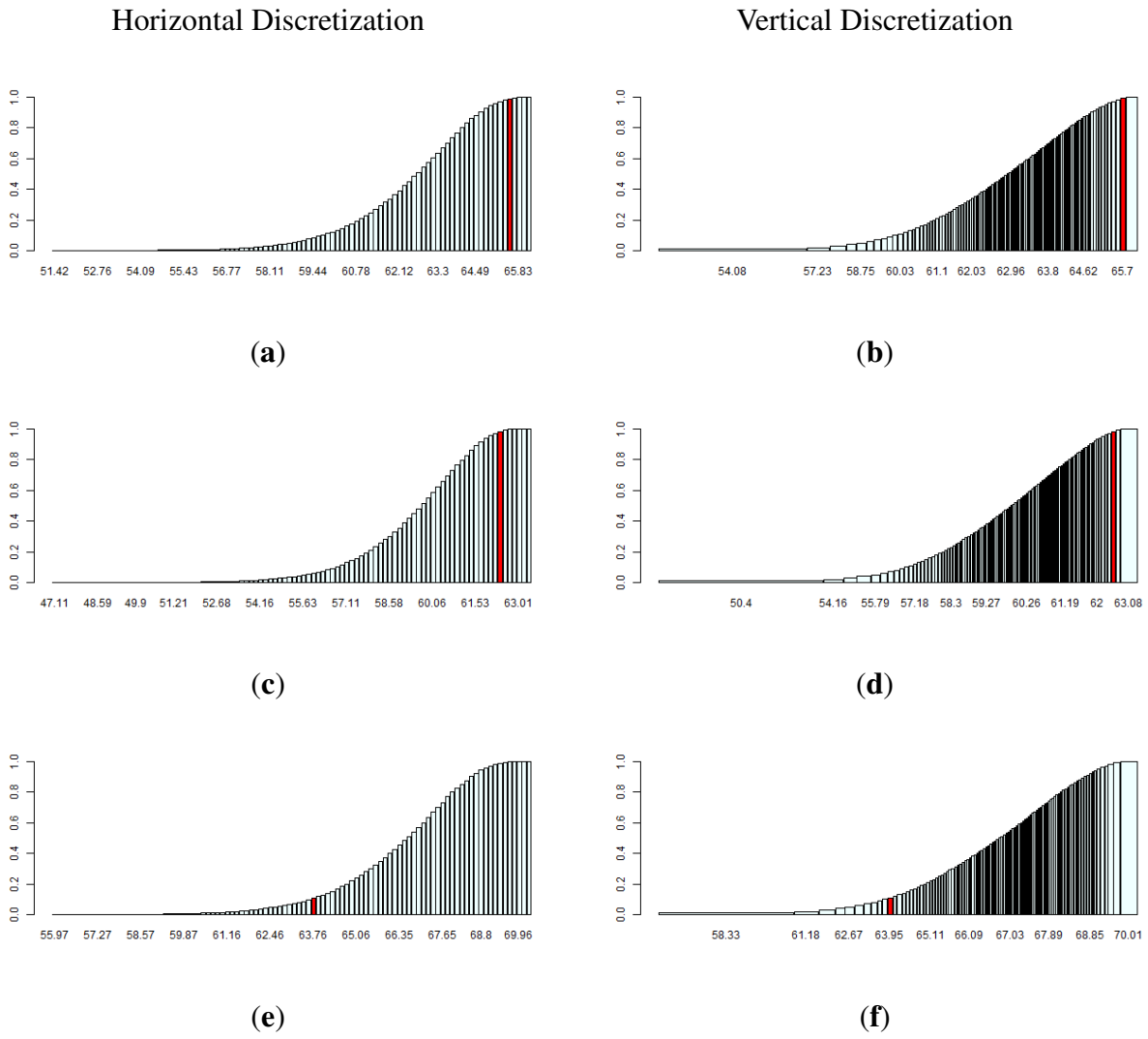
$$Ev_{M_1}(H_1) = 0.9878, Ev_{M_1}(H_2) = 0.9806 \text{ and } Ev_{M_1}(H_3) = 0.1066\,;$$
$$Ev_{M_2}(H_1) = 0.0004, Ev_{M_2}(H_2) = 0.0006 \text{ and } Ev_{M_2}(H_3) = 0.0004,$$

E-values found using vertical discretization:

$$Ev_{M_1}(H_1) = 0.99, Ev_{M_1}(H_2) = 0.98 \text{ and } Ev_{M_1}(H_3) = 0.11\ ;$$
$$Ev_{M_2}(H_1) = 0.01, Ev_{M_2}(H_2) = 0.01 \text{ and } Ev_{M_2}(H_3) = 0.01.$$

Figure 5 shows the histogram of the truth functions, W_1, W_2 and W_3, for the model, M_1 (Y and Z are conditionally independent, given X). In Figure 5a,c,e, 100 bins are uniformly distributed over the x-axis (using the empirical values of $\min f_n(\theta_x)$ and $\max f_n(\theta_x)$). In Figure 5b,d,f, 100 bins are uniformly distributed over the y-axis (each bin represents an increase in 1% in density from the previous bin). The function, W_x, evaluated at the maximum posterior density over the respective hypothesis, $f_n(\theta_x^*)$, in red, corresponds to the e-values found (e.g., $W_3(f(\theta_3^*)) \approx 0.1066$, for the horizontal discretization in Figure 5e).

Figure 5. Histogram with 100 bins for the truth functions of the model, M_1 (Figure 4a for each value of X. (**a**) W_1 for Model M_1, $f_n(\theta_1^*)$ in red; (**b**) W_1 for Model M_1, $f_n(\theta_1^*)$ in red; (**c**) W_2, for Model M_1, $f_n(\theta_2^*)$ in red; (**d**) W_2, for Model M_1, $f_n(\theta_2^*)$ in red; (**e**) W_3, for Model M_1, $f_n(\theta_3^*)$ in red; (**f**) W_3, for Model M_1, $f_n(\theta_3^*)$ in red. In red is the maximum posterior density under the respective elementary component (H_1, H_2 and H_3) of the hypothesis of conditional independence H for both horizontal and vertical discretization procedures.



The evidence supporting the hypothesis of the conditional independence H, as in Equation (31), for each model is as follows:

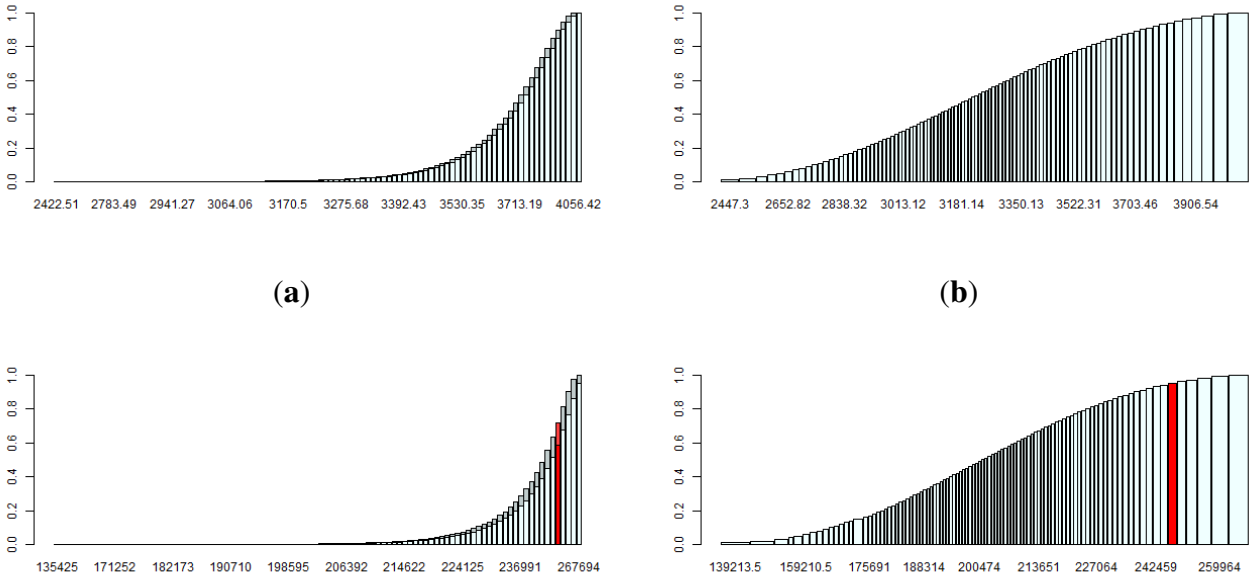
$$Ev(H) = W_1 \otimes W_2 \otimes W_3 \left(f_n(\theta_1^*) \cdot f_n(\theta_2^*) \cdot f_n(\theta_3^*)\right). \tag{37}$$

The convolution follows the commutative property, and the order of the convolutions is therefore irrelevant.

$$W_1 \otimes W_2 \otimes W_3(f) = W_3 \otimes W_2 \otimes W_1(f). \tag{38}$$

Using the algorithm for numerical convolution described in Algorithm 1, we found the convolution of the truth functions, W_1 and W_2, resulting in a cumulative function (W_{12}) with $10,000$ bins (100^2 bins). We then performed the condensation procedures described in Algorithms 2 and 3 and reduced the cumulative distribution to 100 bins, with lower and upper bounds (W_{12}^l and W_{12}^u) for the horizontal condensation. The results are shown in Figure 6a,b for Model M_1 (horizontal and vertical condensations, respectively) and in Figure 7a,b for Model M_2.

Figure 6. Histogram with 100 bins resulting from the convolutions for Model M_1: (**a**) $W_1 \otimes W_2$ with horizontal discretization; (**b**) $W_1 \otimes W_2$ with vertical discretization; (**c**) $W_1 \otimes W_2 \otimes W_3$ with horizontal discretization; (**d**) $W_1 \otimes W_2 \otimes W_3$ with vertical discretization. In red in (**c**) and (**d**) is the bin representing the product of the maximum posterior density under the elementary components (H_1, H_2 and H_3) of the hypothesis of the conditional independence H for model M_1.



The convolution of W_{12} and W_3 was followed by their condensation. The results are shown in Figure 6c,d (Model M_1) and Figure 7c,d (Model M_2).

The e-values supporting the hypothesis of conditional independence for both models are given below.

The intervals for the e-values were found using horizontal discretization and condensation, as follows:

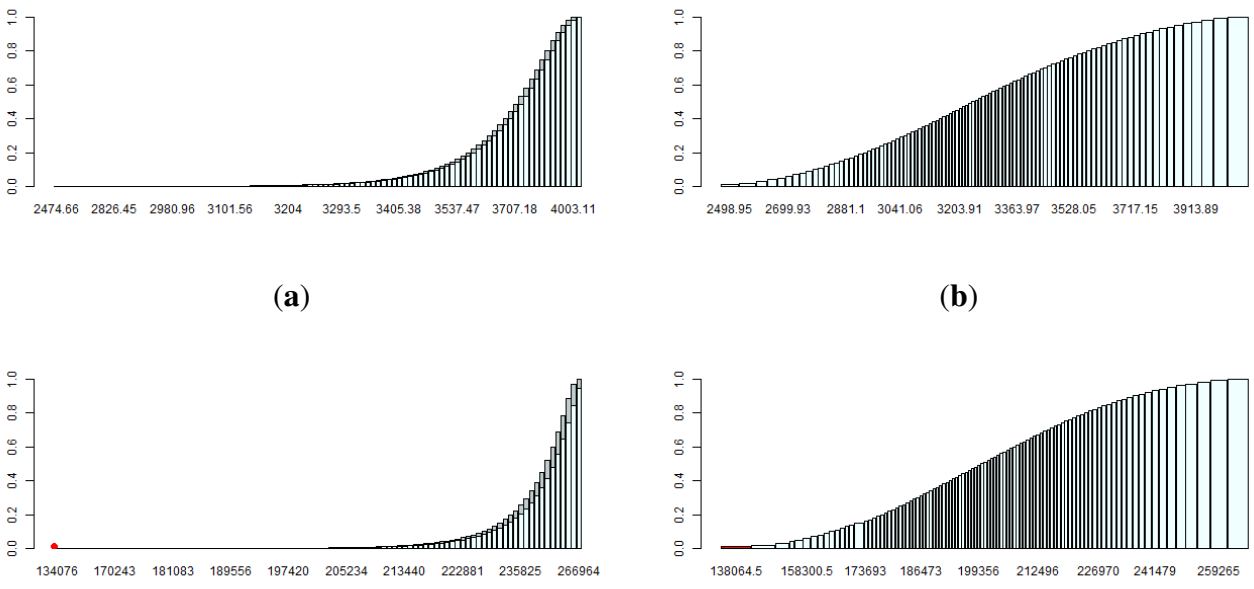
$$Ev_{M_1}(H) = [0.587427, 0.718561] \,,$$
$$Ev_{M_2}(H) = [8 \cdot 10^{-12}, 6.416 \cdot 10^{-9}].$$

The e-values found using vertical discretization and condensation were as follows:

$$Ev_{M_1}(H) = 0.95 \,,$$
$$Ev_{M_2}(H) = 0.01.$$

These results show strong evidence supporting the hypothesis of conditional independence between Y and Z, given X, for Model M_1 (using both discretization/condensation procedures). No evidence supporting the same hypothesis for the second model was found. This result is very relevant and promising as a motivation for further studies on the use of FBST as a CI test for the structural learning of graphical models.

Figure 7. Histogram with 100 bins resulting from the convolutions for model M_2: (**a**) $W_1 \otimes W_2$ with horizontal discretization; (**b**) $W_1 \otimes W_2$ with vertical discretization; (**c**) $W_1 \otimes W_2 \otimes W_3$ with horizontal discretization; (**d**) $W_1 \otimes W_2 \otimes W_3$ with vertical discretization. In red in (**c**) and (**d**) is the bin representing the product of the maximum posterior density under the elementary components (H_1, H_2 and H_3) of the hypothesis of conditional independence, H, for model M_2.



(**a**) (**b**)

(**c**) (**d**)

5. Conclusions and Future Work

This paper provides a framework for performing tests of conditional independence for discrete datasets using the Full Bayesian Significance Test. A simple application of this test includes examining

the structure of a directed acyclic graph given two different models. The result found in this paper suggests that FBST should be considered a good alternative to performing CI tests to uncover the structures of probabilistic graphical models from data.

Future research should include the use of FBST in algorithms to learn the structures of graphs with larger numbers of variables; to increase the capacity for performing these mathematical methods to calculate *e-values* (because learning DAG structures from data requires an exponential number of CI tests to be performed, each CI test needs to be performed faster); and to empirically evaluate the threshold for *e-values* to define conditional independence versus dependence. The last of these areas of future exploration should be achieved by minimizing the linear combination of type I and II errors (incorrect rejection of a true hypothesis of conditional independence and failure to reject a false hypothesis of conditional independence).

Acknowledgment

The authors are grateful for the support of IME-USP, to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Apoio à Pesquisa do Estado de São Paulo (FAPESP).

Author Contribution

All authors made substantial contributions to conception and design, acquisition of data and 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 and any revised version.

Conflicts of Interest

We certify that there is no conflict of interest regarding the material discussed in the manuscript.

References

1. Barlow, R.E.; Pereira, C.A.B. Conditional independence and probabilistic influence diagrams. In *Topics in Statistical Dependence*; Block, H.W., Sampson, A.R., Savits, T.H., Eds.; Lecture Notes-Monograph Series; Institute of Mathematical Statistics: Beachwood, OH, USA, 1990; pp. 19–33.
2. Basu, D.; Pereira, C. A. Conditional independence in statistics. *Sankhy: The Indian Journal of Statistics.* **1983**, *Series A*, 371–384.
3. Pearl, J.; Verma, T.S. *A Theory of Inferred Causation*; Studies in Logic and the Foundations of Mathematics; Morgan Kaufmann: San Mateo, CA, USA, 1995; pp. 789–811.
4. Cheng, J.; Bell, D.A.; Liu, W. Learning belief networks from data: An information theory based approach. In Proceedings of the sixth International Conference on Information and Knowledge Management, Las Vegas, NV, USA, 10–14 November 1997; pp. 325–331.

5. Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **2006**, *65*, 31–78.
6. Yehezkel, R.; Lerner, B. Bayesian network structure learning by recursive autonomy identification. *J. Mach. Learn. Res.* **2009**, *10*, 1527–1570.
7. Pereira, C.A.B.; Stern, J.M. Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy* **1999**, *1*, 99–110.
8. Borges, W.; Stern, J.M. The rules of logic composition for the Bayesian epistemic e-values. *Log. J. IGPL* **2007**, *15*, 401–420.
9. Geenens, G.; Simar, L. Nonparametric tests for conditional independence in two-way contingency tables. *J. Multivar. Anal.* **2010**, *101*, 765–788.
10. Kilicman, A.; Arin, M.R.K. A note on the convolution in the Mellin sense with generalized functions. *Bull. Malays. Math. Sci. Soc.* **2002**, *25*, 93–100.
11. Springer, M.D. *The Algebra of Random Variables*; Wiley: New York, NY, USA, 1979.
12. Williamson, R.C. Probabilistic Arithmetic. Ph.D Thesis, University of Queensland, Austrilia, 1989.
13. Kaplan, S.; Lin, J.C. An improved condensation procedure in discrete probability distribution calculations. *Risk Anal.* **1987**, *7*, 15–19.
14. Williamson, R.C.; Downs, T. Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *Int. J. Approx. Reason.* **1990**, *4*, 89–158.

Appendix

Table A1. Conditional probability distribution tables. (**a**) The distribution of X, (**b**) the conditional distribution of Y, given X, and (**c**) the conditional distribution of Z, given X.

(a) CPT of X

X	p(X)
1	0.3
2	0.2
3	0.5

(b) CPT of Y given X

Y	p(Y|X=1)	p(Y|X=2)	p(Y|X=3)
1	0.3	0.4	0.2
2	0.2	0.4	0.1
3	0.5	0.2	0.7

(c) CPT of Z given X

Z	p(Z|X=1)	p(Z|X=2)	p(Z|X=3)
1	0.5	0.1	0.6
2	0.4	0.1	0.1
3	0.1	0.8	0.3

Table A2. Conditional probability distribution table of Z, given X & Y.

Z	p(Z|X=1,Y=1)	p(Z|X=1,Y=2)	p(Z|X=1,Y=3)
1	0.5	0.1	0.6
2	0.4	0.1	0.1
3	0.1	0.8	0.3
Z	p(Z|X=2,Y=1)	p(Z|X=2,Y=2)	p(Z|X=2,Y=3)
1	0.2	0.4	0.8
2	0.2	0.3	0.1
3	0.6	0.3	0.1
Z	p(Z|X=3,Y=1)	p(Z|X=3,Y=2)	p(Z|X=3,Y=3)
1	0.1	0.5	0.2
2	0.2	0.4	0.6
3	0.7	0.1	0.2

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).