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, \((\Xi, \Delta, \Theta)\), where \(\Xi\) is the sample space, \(\Delta\), the family of measurable subsets of \(\Xi\) and \(\Theta\) the parameter space (\(\Theta\) is a subset of \(\mathbb{R}^n\)).

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

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

The credibility of \(T_\varphi\) 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.

Defining the maximum of the posterior density over the null hypothesis as \(f_x^*\), with the maximum point at \(\theta_0^*\),

\[
\theta_0^* \in \arg\max_{\theta \in \Theta_0} f_x(\theta), \text{ and } f_x^* = f_x(\theta^*), \tag{3}
\]

and defining \(T^* = T_{f_x^*}\) as the tangent set to the null hypothesis, \(H_0\), the credibility of \(T^*\) is \(\kappa^*\).

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 below, where \(\mu\) is the mean of a normal distribution and \(\tau\) 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, $\theta^*_0$, 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:

$$H : \theta_1 \in \Theta_1^H \land \theta_2 \in \Theta_2^H \land \ldots \land \theta_k \in \Theta_k^H,$$

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

$$\theta^*_j \in \arg\max_{\theta_j \in \Theta^H_j} f_x(\theta_j) \quad \text{and} \quad f^*_{j} = f_x(\theta^*_j).$$

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, $v$:

$$R_j(v) = \{ \theta_j \in \Theta_j | f_x(\theta_j) \leq v \},$$

$$W_j(v) = \int_{R_j(v)} f_x(\theta_j) \, d\theta_j. \quad (8)$$

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

$$Ev(H_j) = W_j(f^*_j). \quad (9)$$

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:

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

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

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

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].

$$f_Z(z) = (f_X * f_Y)(z) = \int_0^\infty (1/y) f_X(z/y) f_Y(y) \, dy. \quad (12)$$
The evidence supporting the complex hypothesis can be then described as the Mellin convolution of the truth function of its components:

\[ Ev(H) = W_1 \otimes W_2 \otimes W_3 \otimes \ldots \otimes W_k \left( f_1^* \cdot f_2^* \cdot f_3^* \cdot \ldots \cdot f_k^* \right). \]  

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

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

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,

\[ (W_1 \otimes W_2) \otimes W_3 = W_1 \otimes (W_2 \otimes W_3) = (W_1 \otimes W_3) \otimes W_2 = W_1 \otimes (W_3 \otimes W_2). \]  

### 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:

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

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

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

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 \( X_2 \):

\[ \ln (Y_1) = X_1 \sim \mathcal{N} \left( \mu_1, \sigma_1^2 \right), \]
\[ \ln (Y_2) = X_2 \sim \mathcal{N} \left( \mu_2, \sigma_2^2 \right). \]  

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:

\[ Y_1 = e^{X_1} \text{ and } Y_2 = e^{X_2}, \]
\[ Y_1 \cdot Y_2 = e^{X_1+X_2}, \]
\[ \ln (Y_1 \cdot Y_2) = X_1 + X_2 \sim \mathcal{N} \left( \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 \right), \]
\[ \therefore Y_1 \cdot Y_2 \sim \ln \mathcal{N} \left( \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 \right). \]  

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

\[ W_{12}(y_{12}) = \int_{-\infty}^{y_{12}} f_{Y_1, Y_2} (t) dt, \]

where \( f_{Y_1, Y_2} \) 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 \( F_{X_2} \):

\[
F_Z(z) = \int_0^z F_{X_1}(\frac{z}{t}) \, dF_{X_2}(t). \tag{21}
\]

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.

1: procedure CONVOLUTION($Y_1, Y_2, f_{Y_1}, f_{Y_2}$)  
   \(\triangleright\) Discrete pdf of $Y_1$ and $Y_2$
2: \(f \leftarrow \text{array}(0, \text{size} \leftarrow n^2)\)  
   \(\triangleright\) $f$ and $W$ has $n^2$ bins
3: \(W \leftarrow \text{array}(0, \text{size} \leftarrow n^2)\)
4: \(y_1y_2 \leftarrow \text{array}(0, \text{size} \leftarrow n^2)\)  
   \(\triangleright\) keep $Y_1 \ast Y_2$
5: for $i \leftarrow 1, n$ do
6:   for $j \leftarrow 1, n$ do
7:     \(f[(i - 1) \cdot n + j] \leftarrow f_{Y_1}[i] \cdot f_{Y_2}[j]\)
8:     \(y_1y_2[(i - 1) \cdot n + j] \leftarrow Y_1[i] \ast Y_2[j]\)
9:   end for
10: end for
11: \(\text{sortedIdx} \leftarrow \text{order}(y_1y2)\)  
   \(\triangleright\) find order of $Y_1 \ast Y_2$
12: \(f \leftarrow f[\text{sortedIdx}]\)  
   \(\triangleright\) sort $f$ according to $Y_1 \ast Y_2$
13: \(W[1] \leftarrow f[1]\)
14: for $i \leftarrow k, n^2$ do
15:   \(W[i] \leftarrow W[k] + W[k - 1]\)
16: end for
17: return $W$  
   \(\triangleright\) Discrete cdf of $Y_1 \ast Y_2$
18: end procedure

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

1: procedure HORIZONTALCONDENSATION($W$)  
   \(\triangleright\) Histogram of a cdf with $n^2$ bins
2: \(W^l \leftarrow \text{array}(0, \text{size} \leftarrow n)\)
3: \(W^u \leftarrow \text{array}(0, \text{size} \leftarrow n)\)
4: for $i \leftarrow 1, n$ do
5:   \(W^l[i] \leftarrow W[(i - 1) \cdot n + 1]\)  
   \(\triangleright\) lower bound after condensation
6:   \(W^u[i] \leftarrow W[i \cdot n]\)  
   \(\triangleright\) upper bound after condensation
7: end for
8: return \([W^l, W^u]\)  
   \(\triangleright\) Histograms with upper/lower bounds
9: 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.

1: procedure VERTICALCONDENSATION(W, f, x)  \(\triangleright\) Histograms of a cdf and pdf, and breaks in the x-axis.
2:     breaks ← \([1/n, 2/n, \ldots, 1]\)  \(\triangleright\) uniform breaks in y-axis
3:     \(W_n \leftarrow \text{array}(0, \text{size} \leftarrow n)\)
4:     \(x_n \leftarrow \text{array}(0, \text{size} \leftarrow n)\)
5:     lastbreak ← 1
6:     \(i \leftarrow 1\)
7:     for all \(b \in \text{breaks}\) do
8:         \(w \leftarrow \text{first}(W \geq b)\)  \(\triangleright\) find break to create current bin
9:         if \(W[w] \neq b\) then  \(\triangleright\) if the break is within a current bin
10:             ratio ← \((b - W[w - 1])/(W[w] - W[w - 1])\)
11:             \(x_n[i] \leftarrow \frac{1}{1/n} \left(\text{sum}(f[w - 1] \cdot x[w - 1]) + \text{ratio} \cdot f[w] \cdot x[w]\right)\)
12:             \(W[i - 1] \leftarrow b\)
13:             \(W_n[i] \leftarrow b\)
14:             \(f[i - 1] \leftarrow f[w - 1] + \text{ratio} \cdot f[w]\)
15:             \(f[i] \leftarrow (1 - \text{ratio}) \cdot f[w]\)
16:         else
17:             \(x_n[i] \leftarrow x[w]\)
18:             \(W_n[i] \leftarrow W[w]\)
19:         end if
20:     lastbreak ← b
21:     \(i \leftarrow i + 1\)
22: end for
23: return \([W_n, x_n]\)  \(\triangleright\) Histograms with upper/lower bounds
24: 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.

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.
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 $Y \perp Z|X$ can be written as the complex null hypothesis, $H$:

$$H : [Y \perp Z|X = 1] \land [Y \perp Z|X = 2] \land \cdots \land [Y \perp Z|X = k]. \quad (22)$$

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

$$H_1 : Y \perp Z|X = 1$$
$$H_2 : Y \perp Z|X = 2$$
$$\cdots$$
$$H_k : Y \perp Z|X = k$$

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 $H_x$); $n_{yxz}$ is the count of $[Y,Z] = [y,z]$ when $X = x$.

<table>
<thead>
<tr>
<th></th>
<th>$Z = 1$</th>
<th>$Z = 2$</th>
<th>$\cdots$</th>
<th>$Z = c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 1$</td>
<td>$n_{11x}$</td>
<td>$n_{12x}$</td>
<td>$\cdots$</td>
<td>$n_{1cx}$</td>
</tr>
<tr>
<td>$Y = 2$</td>
<td>$n_{21x}$</td>
<td>$n_{22x}$</td>
<td>$\cdots$</td>
<td>$n_{2cx}$</td>
</tr>
<tr>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$Y = r$</td>
<td>$n_{r1x}$</td>
<td>$n_{r2x}$</td>
<td>$\cdots$</td>
<td>$n_{rcx}$</td>
</tr>
</tbody>
</table>

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 $\theta_x = [\theta_{y1z}, \theta_{y2z}, \ldots, \theta_{ycz}]$.

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

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

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:

$$f(n_x|\theta_x) = n_{..x}! \prod_{y,z}^{r,c} \frac{\theta_{yzzx}^{n_{yzx}}}{n_{yzx}!}. \quad (24)$$
The posterior distribution is thus a Dirichlet distribution, $f_n(\theta_x)$:

$$f_n(\theta_x) \propto \prod_{y,z} \theta_{yzx}^{n_{yzx} + \alpha_{yzx} - 1}. \quad (25)$$

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

$$H_x : \theta_{yzx} = \theta_{y.x} \cdot \theta_{z.x}, \forall y, z \quad (26)$$

The elementary components of hypothesis $H$ are as follows:

$$H_1 : \theta_{yz1} = \theta_{z1} \cdot \theta_{y.1}, \forall y, z$$
$$H_2 : \theta_{yz2} = \theta_{z2} \cdot \theta_{y.2}, \forall y, z$$
$$\ldots$$
$$H_k : \theta_{yzk} = \theta_{z.k} \cdot \theta_{y.k}, \forall y, z \quad (27)$$

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, $\theta_x$). The maximum density ($f^*_x$) is the posterior density evaluated at this point.

$$\theta_x^{*yz} = \frac{n_{H_x}^{yz} + \alpha_{yzx} - 1}{n_{r,c}^{H_x} + \alpha_{y.x} - r \cdot c} \text{ and } f^*_x = f_n(\theta_x^*) \quad (28)$$

where $\theta_x^* = [\theta_{11x}^*, \ldots, \theta_{r,cx}^*]$. The evidence supporting $H_x$ can be written in terms of the truth function, $W_x$, as defined in Section 3:

$$R_x(f) = \{ \theta_x \in \Theta_x | f_x(\theta_x) \leq f \},$$
$$W_x(f) = \int_{R_x(f)} f_n(\theta_x) d\theta_x \propto \int_{R_x(f)} \prod_{y,z} \theta_{yzx}^{n_{yzx} + \alpha_{yzx} - 1} d\theta_x. \quad (29)$$

The evidence supporting $H_x$ is:

$$Ev(H_x) = W_x(f^*_x). \quad (30)$$

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 $H$:

$$Ev(H) = W_1 \otimes W_2 \otimes \ldots \otimes W_k (f^*_1 \cdot f^*_2 \cdot \ldots \cdot f^*_k). \quad (31)$$

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 $X$.

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:

\[
\begin{align*}
H_1 : Y \perp \!\!\!\perp Z | X &= 1 \\
H_2 : Y \perp \!\!\!\perp Z | X &= 2 \\
H_3 : Y \perp \!\!\!\perp Z | X &= 3
\end{align*}
\]

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)):

\[
\alpha_1 = \alpha_2 = \alpha_3 = [1, 1, 1], \\
f(\theta_1 | \alpha_1) = f(\theta_3 | \alpha_3) = f(\theta_4 | \alpha_3) = 1. \tag{32}
\]

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:

\[
f_n(\theta_2) \propto \theta_{112}^{42} \cdot \theta_{122}^{14} \cdot \theta_{132}^{32} \cdot \theta_{212}^{15} \cdot \theta_{222}^{23} \cdot \theta_{232}^{31} \cdot \theta_{312}^{11} \cdot \theta_{322}^{21} \cdot \theta_{332}^{32}. \tag{34}
\]
The point of highest density, in this example, following the hypothesis of independence (Equations (26) and (28)), was found to be the following:

$$\theta^*_2 \approx [0.036, 0.039, 0.317, 0.038, 0.041, 0.329, 0.019, 0.020, 0.162].$$  

(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:

$$R_2(f) = \{ \theta_2 \in \Theta_2 \mid f_n(\theta_2) \leq f \} ,$$

$$W_2(f) = \int_{R_2(f)} f_n(\theta_2) \, d\theta_2 ,$$

$$Ev_{M_1}(H_2) = W_2(f_n(\theta^*_2)).$$  

(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.

<table>
<thead>
<tr>
<th></th>
<th>(a) Model $M_1$ (for $X = 1$)</th>
<th>(b) Model $M_2$ (for $X = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Z = 1$</td>
<td>$Z = 2$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>241</td>
<td>187</td>
</tr>
<tr>
<td>$Y = 2$</td>
<td>139</td>
<td>130</td>
</tr>
<tr>
<td>$Y = 3$</td>
<td>364</td>
<td>302</td>
</tr>
<tr>
<td></td>
<td>744</td>
<td>619</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(c) Model $M_1$ (for $X = 2$)</th>
<th>(d) Model $M_2$ (for $X = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Z = 1$</td>
<td>$Z = 2$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>$Y = 2$</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>$Y = 3$</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(e) Model $M_1$ (for $X = 3$)</th>
<th>(f) Model $M_2$ (for $X = 3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Z = 1$</td>
<td>$Z = 2$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>282</td>
<td>35</td>
</tr>
<tr>
<td>$Y = 2$</td>
<td>131</td>
<td>37</td>
</tr>
<tr>
<td>$Y = 3$</td>
<td>1,055</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>1,468</td>
<td>215</td>
</tr>
</tbody>
</table>

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:

$$Ev(H) = W_1 \otimes W_2 \otimes W_3 (f_n(\theta_1^*) \cdot f_n(\theta_2^*) \cdot f_n(\theta_3^*)).$$

(37)

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

$$W_1 \otimes W_2 \otimes W_3(f) = W_3 \otimes W_2 \otimes W_1(f).$$

(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 ($W_{12l}$ and $W_{12u}$) 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 \(M_2\).

### 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


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 $X$.

<table>
<thead>
<tr>
<th>(a) CPT of $X$</th>
<th>(b) CPT of $Y$ given $X$</th>
<th>(c) CPT of $Z$ given $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$Y$</td>
<td>$Z$</td>
</tr>
<tr>
<td>$p(X)$</td>
<td>$p(Y</td>
<td>X=1)$ $p(Y</td>
</tr>
<tr>
<td>1 0.3</td>
<td>1 0.3 0.2 0.4 0.2</td>
<td>1 0.5 0.1 0.6</td>
</tr>
<tr>
<td>2 0.2</td>
<td>2 0.2 0.4 0.1</td>
<td>2 0.4 0.1</td>
</tr>
<tr>
<td>3 0.5</td>
<td>3 0.5 0.2 0.7</td>
<td>3 0.1 0.8</td>
</tr>
</tbody>
</table>
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/).