Bayesian test of significance for conditional independence: The multinomial model

Pablo M. Andrade[*], Julio M. 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

June 18, 2013

Abstract

Conditional independence tests (CI 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 (PGM)–which includes Bayesian Networks (BN) models–CI tests are especially important for the task of learning the PGM structure from data. In this paper, we propose the Full Bayesian Significance Test (FBST) for tests of conditional independence for discrete datasets. FBST is a powerful Bayesian test for precise hypothesis, as an alternative to frequentist's significance tests (characterized by the calculation of the *p-value*).

1 Introduction

Barlow and Pereira (1990) discuss a graphical approach to conditional independence. A probabilistic influence diagram is a directed acyclic graph (DAG) that helps to model statistical problems. The graph is composed of

[*]Email: pablo.andrade@usp.br
[†]Email: jstern@ime.usp.br
[‡]Email: cpereira@ime.usp.br

a set of nodes or vertices, representing the variables, and a set of arcs joining the nodes, representing the dependence relationships shared by these variables.

The construction of the model helps to understand the problem and gives a good representation of interdependence of the variables involved in the problem. The joint probability of these variable can be written as a product of conditional distributions, based on the relationships of independence and conditional independence among the variables involved in the problem.

Sometimes the interdependence of the variables is not known, and in this case, the model structure is required to be learnt from data. Algorithms such as the *IC-Algorithm (Inferred Causation)* described in Pearl and Verma (1995) are designed to uncover these structures from data. This algorithm uses a series of CI tests to remove and direct the arcs connecting the variables in the model, returning a DAG that minimally (with the minimum number of parameters, without loss of information) represents the variables in the problem.

The problem of learning DAG structures from data motivates the proposal of new powerful statistical tests for the hypothesis of conditional independence, since the accuracy of structures learnt are directly affected by errors committed by these tests. Recently proposed structure learning algorithms (see Cheng et al., 1997; Tsamardinos et al., 1997; Yehezkel and Lerner, 2009) indicate as main source of errors the results of CI tests.

In this paper, we propose the Full Bayesian Significance Test (FBST) for tests of conditional independence for discrete datasets. FBST is a powerful Bayesian test for precise hypothesis, and can be used to learn DAG structures from data, as an alternative to CI test currently used, such as *Pearson's* χ^2 *test*.

This paper is organized as follows. In Section 2 we review the Full Bayesian Significance Test (FBST). In Section 3, we review the FBST for composite hypothesis. Section 4 shows an example of test of conditional independence used to learn a simple model with 3 variables.

2 The Full Bayesian Significance Test

The Full Bayesian Significance Test (FBST) is presented by Pereira and Stern (1999) as a coherent Bayesian significance test for sharp hypothesis. 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, where the posterior density of each of its elements is greater than the maximum of the posterior density over the Null hypothesis. 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.

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

The credibility of T_φ is given by its posterior probability:

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

, 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 maximum point at θ_0^*:

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

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

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

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

If the probability of the set T^* is large, the null set is in a region of low probability and the evidence is against the Null hypothesis H_0. But, 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

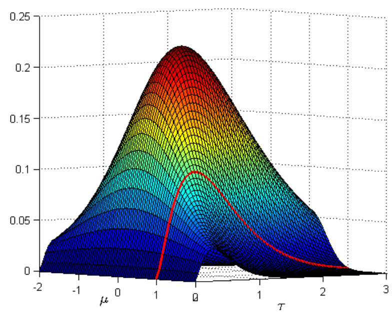
Figure 1 shows the tangent set for a Null hypothesis $H_0 : \mu = 1$, for the posterior distribution f_x given bellow, where μ is the mean of a normal distribution and τ, the *precision* (the inverse of the variance $\tau = \frac{1}{\sigma^2}$):

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

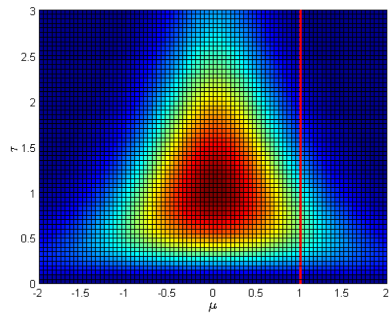
3 FBST: Compositionality

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 (FBST), is analysed in Borges and Stern (2007).

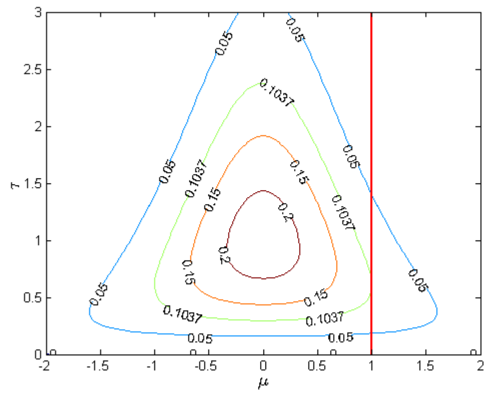
For a given set of *independent* parameters $(\theta_1, \ldots, \theta_k) \in (\Theta_1 \times \ldots \times \Theta_k)$, a complex hypothesis H, such as:



(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: $\mu = 1.0$.

Figure 1: Example of tangent set for a 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 density 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 density greater than f^*), and the e-value of H_0 is the complement of the integral of f_x bounded by the green contour line.

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

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

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

, and the credibility of H can be evaluated based on the credibility of these components. The evidence in favour of the complex hypothesis H (measured by its *e-value*) can not be obtained directly from the evidence in favour of the elementary components, but based on their *Truth Function* W^j (or cumulative surprise distribution) 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) constrained to the subset of the parameter space defined by hypothesis H_j:

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

The *truth function* W_j is the probability of the region of the parameter space, where the posterior density is lower or equal than a value f:

$$R_j(f) = \{\theta_j \in \Theta_j | f_x(\theta_j) \leq f\}$$
$$W_j(f) = \int_{R_j(f)} f_x(\theta_j)\, d\theta_j$$

And the evidence supporting the hypothesis H_j is:

$$Ev(H_j) = W_j(f_j^*)$$

The evidence supporting the complex hypothesis can be then described in terms of the *truth function* of its components, as the Mellin convolution of these functions:

$$Ev(H) = W_1 \otimes W_2 \otimes W_3 \otimes \ldots \otimes W_k\,(f_1^* \cdot f_2^* \cdot f_3^* \cdot \ldots \cdot f_k^*)$$

Where the Mellin Convolution of two *truth functions*, $W_1 \otimes W_2$, is the distribution function:

$$W_1 \otimes W_2(x) = \int_0^x W_1\left(\frac{x}{y}\right) W_2(y) dy$$

3.1 Numerical Method for Convolution and Condensation

Williamson and Downs (1990) investigate numerical procedures to handle arithmetic operations for random variables. Replacing basic operations of arithmetic, used for fixed numbers, by convolutions, they show how to calculate the joint distribution for a set of random variables and their respective upper and lower bounds.

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}:

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

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 in Williamson and Downs 1990, page 188).

The numerical convolution of two distributions with N bins returns a distribution with N^2 bins. For a sequence of operations, this would be a problem, since the result of each operation would be larger than the input for the operations. The authors, hence, propose a simple method to reduce the size of the output to N bins, without introducing error to the result. This operation is called *condensation*, and it 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.

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

procedure CONVOLUTION(f_{Y_1}, f_{Y_2}) ▷ Discrete p.d.f. of Y_1 and Y_2
  $f \leftarrow array(0, size \leftarrow n^2)$ ▷ f and W has n^2 bins
  $W \leftarrow array(0, size \leftarrow n^2)$
  for $i \leftarrow 1, n$ **do** ▷ f_1 and f_2 have n bins
    for $j \leftarrow 1, n$ **do**
      $f[(i-1) \cdot n + j] \leftarrow f_{Y_1}[i] \cdot f_{Y_2}[j]$
    end for
  end for
  $W[1] \leftarrow f[1]$
  for $i \leftarrow k, n^2$ **do** ▷ find c.d.f. of $Y_1 \cdot Y_2$
    $W[k] \leftarrow f[k]$
    $W[k] \leftarrow W[k] + W[k-1]$
  end for
  return W ▷ Discrete c.d.f. of $Y_1 \cdot Y_2$
end procedure

Algorithm 2 Find upper lower bound for a c.d.f. for condensation.

procedure HORIZONTALCONDENSATION(W) ▷ Histogram of a c.d.f. 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

3.1.1 Vertical Condensation

Kaplan and Lin (1987) propose a *vertical* condensation procedure for discrete probability calculations, where the condensation is done using the vertical axis, instead of the horizontal axis, as in Williamson and Downs (1990).

The advantage of this approach is that it provides more control over the representation of the distribution, since, 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 of the range of the cumulative distribution in $[0, 1]$ that should be represented by each bin.

In this case, it is also possible to concentrate the attention in a specific region of the distribution. For example, if there is a greater interest in the behaviour 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 convolution followed by condensation procedure, using both approaches is given in Section 3.2. We used, for this example, discretization and condensation procedures with 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 in the interval $[0, 1]$. Algorithm 3 shows the condensation with bins uniformly distributed over the vertical axis.

Algorithm 3 Condensation with bins vertically uniformly distributed.

procedure VERTICALCONDENSATION(W,f,x) ▷ Histograms of a c.d.f. and p.d.f., 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

3.2 Mellin Convolution: Example

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

Assume Y_1 and Y_2, continuous random variables, such that.

$$Y_1 \sim \ln\mathcal{N}\left(\mu_1, \sigma_1^2\right), \text{ and } Y_2 \sim \ln\mathcal{N}\left(\mu_2, \sigma_2^2\right)$$

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

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

, 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:

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

And we can find the distribution of the product of these random variables $(Y_1 \cdot Y_2)$, using simple arithmetic operations, to be also Log-normal:

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

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

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

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

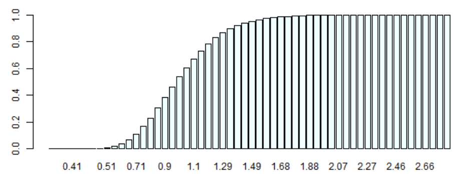
Figure 2 shows the cumulative distribution functions of Y_1 and Y_2 discretized with bins uniformly distributed over both x and y axes (horizontal and vertical discretizations). Figure 3 shows an example of convolution followed by condensation, using both horizontal and vertical condensation procedures,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

4 Test of Conditional Independence in Contingency table using FBST

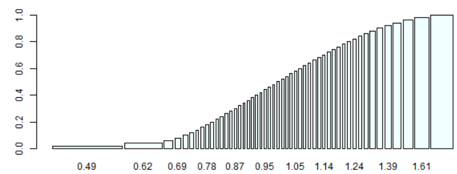
We now apply the methods shown in the previous sections to find the evidence of a complex Null hypothesis of conditional independence, for discrete variables.

Given the discrete random variables X, Y and Z, with X taking values in $\{1, \ldots, k\}$. The test of conditional independence $Y \perp\!\!\!\perp Z|X$ can be written as the complex Null hypothesis H:

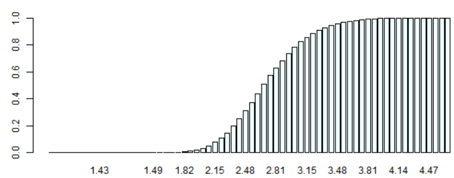
$$H : [Y \perp\!\!\!\perp Z|X = 1] \wedge [Y \perp\!\!\!\perp Z|X = 2] \wedge \cdots \wedge [Y \perp\!\!\!\perp Z|X = k]$$



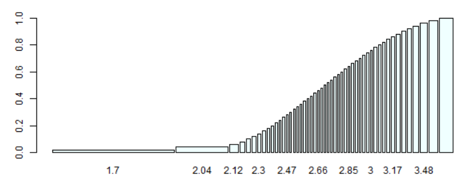
(a) W_1: Horinzontal discretization



(b) W_1: Vertical discretization



(c) W_2: Horinzontal discretization



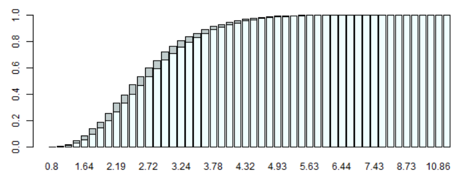
(d) W_2: Vertical discretization

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

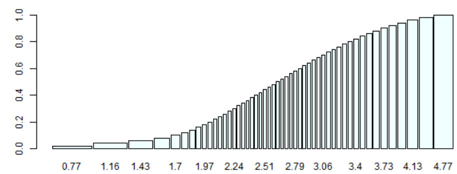
The hypothesis H, can be decomposed in its elementary components:

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

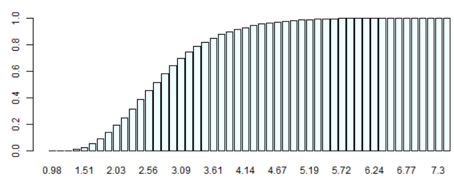
Notice that the hypotheses $H_1, \ldots, H_k$ are *independent*: for each value x taken by X, the values taken by variables Y and Z are assumed to be random observations drawn from some distribution $p(Y, Z|X = x)$. Each of the elementary components is a hypothesis of independence in a contingency table. Table 1 shows the contingency table for Y and Z taking values, respectively, in $\{1, \ldots, r\}$ and $\{1, \ldots, c\}$. 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, the Dirichlet distribution for the vector of parameters $\theta_x = [\theta_{11x}, \theta_{12x}, \ldots, \theta_{rcx}]$.



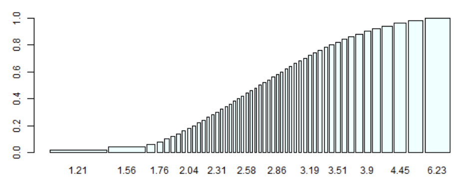
(a) $W_1 \otimes W_2$: Horizontal condensation



(b) $W_1 \otimes W_2$: Vertical condensation



(c) $Y_1 \cdot Y_2$: Horizontal discretizarion



(d) $Y_1 \cdot Y_2$: Vertical discretizarion

Figure 3: Example of convolution of two random variables (Y_1 and Y_2) with Log-normal distribution. The result of the convolution $Y_1 \otimes Y_2$, followed by horizontal condensation (bins uniformly distributed over x-axis) is shown in (a), and by vertical condensation (bins uniformly distributed over 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 horizontal and vertical discretization procedures.

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

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

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:

$$f\left(n_x|\theta_x\right) = n_{..x}! \prod_{y,z}^{r,c} \frac{\theta_{yzx}^{n_{yzx}}}{n_{yzx}!} \tag{2}$$

The posterior distribution will be, then, a Dirichlet distribution $f_n(\theta_x)$:

$$f_n\left(\theta_x\right) \propto \prod_{y,z}^{r,c} \theta_{yzx}^{\alpha_{yzx}+n_{yzx}-1} \tag{3}$$

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}

Under the hypothesis H_x, we have $Y \perp\!\!\!\perp Z|X = x$. In this case, we have that 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 θ_x, in this case, we have:

$$H_x : \theta_{yzx} = \theta_{.zx} \cdot \theta_{y.x}, \forall y, z \tag{4}$$

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

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

$$\theta_{yzx}^* = \frac{n_{yzx}^{H_x} + \alpha_{yzx} - 1}{n_{..x}^{H_x} + \alpha_{..x} - r \cdot c} \text{ and } f_x^* = f_n(\theta_x^*) \tag{5}$$

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

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

$$R_x(f) = \{\theta_x \in \Theta_x | f_x(\theta_x) \leq f\} \tag{6}$$

$$W_x(f) = \int_{R_x(f)} f_n(\theta_x)\, d\theta_x \propto \int_{R_x(f)} \prod_{y,z}^{r,c} \theta_{yzx}^{\alpha_{yzx} + n_{yzx} - 1} d\theta_x \tag{7}$$

And the evidence supporting the hypothesis H_x, is:

$$Ev(H_x) = W_x(f_x^*) \tag{8}$$

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

$$Ev(H) = W_1 \otimes W_2 \otimes \ldots \otimes W_k \left(f_1^* \cdot f_2^* \cdot \ldots \cdot f_k^*\right) \tag{9}$$

The *e-value* for hypothesis H can be found using modern mathematical methods of integration. An example is given in the next section, where the numerical convolution followed by the condensation procedures described in Section 3.1 are used. The application of the method of horizontal condensation results in a 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 CI test using the Full Bayesian Significance Test (FBST) for conditional independence using samples from two different model. For both models, we test if the variable Y is conditionally independent of Z given X.

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

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

The two probabilistic graphical models (M_1 and M_2) are shown in Figure 4, where all 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*, while 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 A.

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

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

Table 2: Contingency tables of Y and Z for a given the value of X for 5,000 random samples. In (a),(c),(e) samples from model M_1 (Figure 4a) for $X = 1$,2 and 3, respectively, in (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	1507

(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	1048

(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	1034

(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	1036

(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
$Y = 3$	1055	143	546	1744
	1468	215	776	2459

(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	1049	372	1726
	464	1240	753	2457

For each model, $5,000$ random observation have been generated, the contingency table of Y and Z for each value of X are shown in Table 2. The hyperparameters of the prior distribution were all set to 1 , the priori is then equivalent to a uniform distribution (from Equation 1):

$$\alpha_1 = \alpha_2 = \alpha_3 = [1, 1, 1]$$
$$f(\theta_1|\alpha_1) = f(\theta_3|\alpha_3) = f(\theta_3|\alpha_3) = 1$$

The posterior distribution, found using Equations 2 and 3, is then:

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

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

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

And the point of highest density, for this example, under the hypothesis of independence (Equations 4 and 5) was found to be:

$$\theta_2^* \approx [0.036, 0.039, 0.317, 0.038, 0.041, 0.329, 0.019, 0.020, 0.162]$$

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 6 and 8, are:

$$R_2(f) = \{\theta_2 \in \Theta_2 | f_n(\theta_2) \leq f\}$$
$$W_2(f) = \int_{R_2(f)} f_n(\theta_2)\, d\theta_2$$
$$Ev_{M_1}(H_2) = W_2(f_n(\theta_2^*))$$

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

E-values found using *horizontal* discretization:

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

, and the *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 Figures 5a, 5c and 5e, 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 Figures 5b, 5d and 5f, 100 bins are uniformly distributed over the y axis (each bin represents an increase in 1% in density from the previous bin). Notice that the functions W_x evaluated at the maximum posterior density over the respective hypothesis $f_n(\theta_x^*)$, in red, correspond to the e-values found (e.g., $W_3(f(\theta_3^*)) \approx 0.1066$, for the horizontal discretization in Figure 5e).

The evidence supporting the hypothesis of conditional independence H, as in Equation 9, for each model, will be:

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

The convolution has commutative property, therefore the order of the convolutions is irrelevant:

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

, 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 3, reducing 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 Figures 6a and 6b for Model M_1 (horizontal and vertical condensations, respectively), and, 7a and 7b for model M_2.

The convolution of W_{12} and W_3 was, then, performed, followed by condensation. The results, are shown in Figures 6c and 6d (model M_1), and 7c and 7d (model M_2).

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

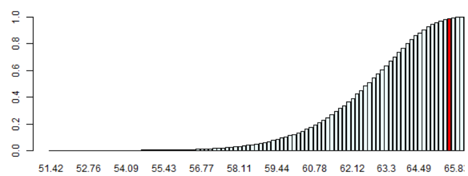
The intervals for the *e-values* found using horizontal discretization and condensation were:

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

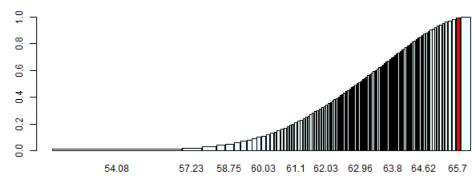
, and the *e-values* found using vertical discretization and condensation were:

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

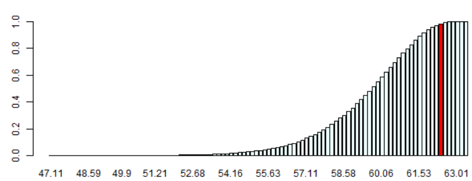
These results show strong evidence supporting the hypothesis of conditional independence between Y and Z given X for the model M_1 (using both discretization/condensation procedures). And no evidence supporting the same hypothesis for the second model. This result is very relevant and promising as a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



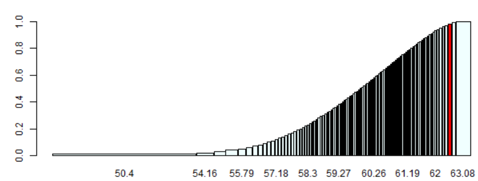
(a) W_1 for Model M_1, $f_n(\theta_1^*)$ in red. Horizontal Discretization.



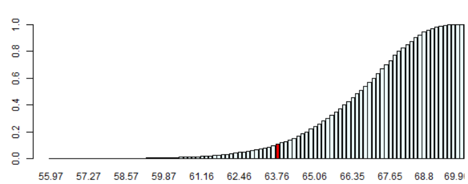
(b) W_1 for Model M_1, $f_n(\theta_1^*)$ in red. Vertical Discretization.



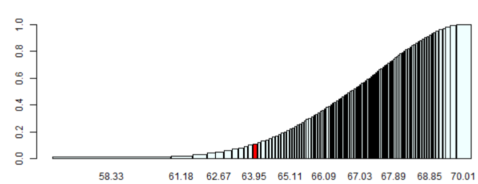
(c) W_2, for Model M_1, $f_n(\theta_2^*)$ in red. Horizontal Discretization.



(d) W_2, for Model M_1, $f_n(\theta_2^*)$ in red. Vertical Discretization.

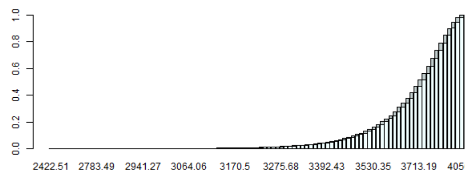


(e) W_3, for Model M_1, $f_n(\theta_3^*)$ in red. Horizontal Discretization.

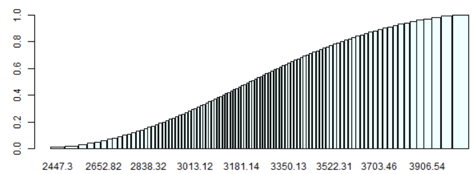


(f) W_3, for Model M_1, $f_n(\theta_3^*)$ in red. Vertical Discretization.

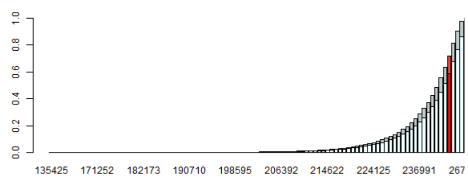
Figure 5: Histogram with 100 bins of the truth functions for the Model M_1 (Figure 4a), for each value of X. In red, the maximum posterior density under the respective elementary component (H_1, H_2 and H_3) of the hypothesis of conditional independence H, for both horizontal and vertical discretization procedures.



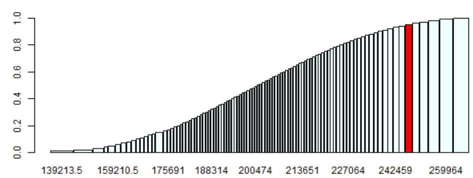
(a) $W_1 \otimes W_2$ for Model M_1.
Horizontal Discretization.



(b) $W_1 \otimes W_2$ for Model M_1.
Vertical Discretization.

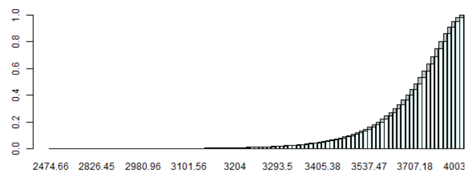


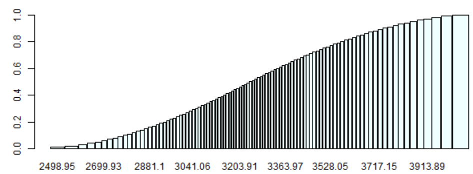
(c) $W_1 \otimes W_2 \otimes W_3$ for Model M_1.
Horizontal Discretization.



(d) $W_1 \otimes W_2 \otimes W_3$ for Model M_1.
Vertical Discretization.

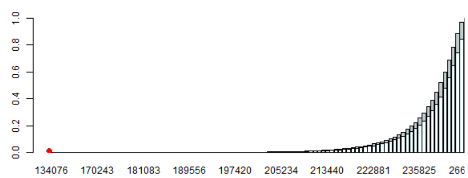
Figure 6: Histogram with 100 bins of the resulting convolutions for Model M_1: (a) $W_1 \otimes W_2$ with horizontal discretization; (b) $W_1 \otimes W_2$ with vertical discretization; (c) $W_1 \otimes W_2 \otimes W_3$ with horizontal discretization; (d) $W_1 \otimes W_2 \otimes W_3$ with vertical discretization.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

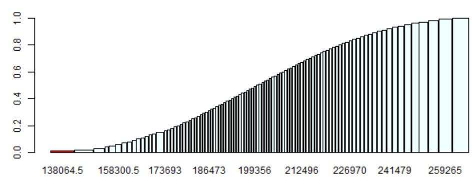




(a) $W_1 \otimes W_2$ for Model M_2.
Horizontal Discretization.

(b) $W_1 \otimes W_2$ for Model M_2.
Vertical Discretization.





(c) $W_1 \otimes W_2 \otimes W_3$ for Model M_2.
Horizontal Discretization.

(d) $W_1 \otimes W_2 \otimes W_3$ for Model M_2.
Vertical Discretization.

Figure 7: Histogram with 100 bins of the resulting 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), the bin representing the product of 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 Conclusion and Future Work

This paper gives the framework to perform tests of conditional independence for discrete datasets using the Full Bayesian Significance Test (FBST). A simple example of application of this test to learn the structure of a directed acyclic graph is given using two different models. The result found in this paper suggests that FBST should be considered as a good alternative to perform CI tests for the task of learning structures of probabilistic graphical models from data.

Future researches include the use of FBST in an algorithm to learn structures of graphs with larger number of variables; the increase in performance of the mathematical methods used to calculate the e-values (as learning DAG structures from data requires an exponential number of CI tests to be performed, each CI test needs to be performed faster); and an empirical evaluation of the threshold for e-values in order to define conditional independence versus dependence, by minimizing a linear combination of errors of type I and II (incorrect rejection of true hypothesis of conditional independence and failure to reject a false hypothesis of conditional independence).

References

Barlow, R. E., & Pereira, C.A.B. (1990). Conditional independence and probabilistic influence diagrams. *Lecture Notes-Monograph Series*, pp. 19–33.

Basu, D., & Pereira, C.A.B. (2011). Conditional independence in statistics. In *Selected Works of Debabrata Basu*, pp. 371–384. Springer New York.

Borges, W., & Stern, J. M. (2007). The rules of logic composition for the Bayesian epistemic e-values. *Logic Journal of IGPL*, **15**(5-6), pp. 401–420.

Cheng, J., Bell, D. A., & Liu, W. (1997, January). Learning belief networks from data: An information theory based approach. In *Proceedings of the sixth international conference on Information and knowledge management*, pp. 325–331. ACM.

Kaplan, S., & Lin, J. C. (1987). An improved condensation procedure in discrete probability distribution calculations. *Risk Analysis*, **7**(1), 15–19.

Pereira, C.A.B., & Stern, J.M. (1999) Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy*, 1, pp. 99-110.

Pearl, J., & Verma, T. S. (1995). A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics*, 134, pp. 789–811.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1), pp. 31–78.

Williamson, R. C. (1989). Probabilistic arithmetic (Doctoral dissertation, University of Queensland).

Williamson, R. C., & Downs, T. (1990) Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2), pp. 89-158.

Yehezkel, R., & Lerner, B. (2009). Bayesian network structure learning by recursive autonomy identification. *The Journal of Machine Learning Research*, 10, pp. 1527–1570.

A Appendix

Table 3: Conditional probability distribution tables. In (a) the distribution of X, in (b) conditional distribution of Y, given X, in (c) 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 4: 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