A New Score for Independence Based on the Imprecise Dirichlet Model

Joaquín Abellán
Department of Computer Science and
Artificial Intelligence.
University of Granada, Spain.
jabellan@decsai.ugr.es

Serafín Moral
Department of Computer Science and
Artificial Intelligence.
University of Granada, Spain.
smc@decsai.ugr.es

Abstract

In this paper we present a new score to determine when two categorical variables are independent. It represents a measure that can be used in classification. It is an interval-valued score that is based on the Heckerman, Geiger, and Chickering's score. We also carry out an empirical comparison with different scores to determine when two binary variables are independent. The others measures that have been considered are: the Bayesian score metric, the Bayesian information criterion (BIC), the p-value of the Chi-square test for independence and the upper entropy score based on imprecise probabilities. For the new score, we find a behaviour that it is more similar to statistical tests from small samples and to Bayesian procedures for large samples. This makes it very appropriate for some concrete types of problems.

Keywords. Independence, statistical tests, Bayesian score, Chi-square test, imprecise Dirichlet model.

1 Introduction

When we have a sample and we want to induce a model from it, one of the main problems is to decide about its complexity. In probabilistic models, the main criterion to determine the final complexity is based on the independence relationships that can be obtained from the sample. Unfortunately, there is not a single criterion to be used when considering whether an independence relationship is supported by the data and none of the existing ones can be considered as superior to the others. All are based on some basic methodology and some additional assumptions or approximations. In this way, we have methods based on frequentist statistical tests of independence [15], on Bayesian scores [5], on the minimum description length principle [14], or on the theory of imprecise probabilities [11]. The quality of a measure will depend on the appropriateness of the assumptions, the error of the approximations, or the generic principle on which it is based, and these aspects are, in general, difficult to assess in a concrete application.

In a previous paper [11] we carried out an empirical evaluation of the different criteria, showing that none of them is superior to the others in any circumstance. Though the Bayesian scores showed a good performance, in general, their behaviour is not satisfactory. Their main problem is that they can decide for dependence when the variables are independent from a sample of very small size. When applied to learning Bayesian networks, we can obtain links that are not supported by the data. As an example, we have considered a Bayesian network with 12 binary independent variables. The marginal distribution for all of them is 0.5 of probability for each one of its values. From it, we have obtained by simulation a sample of size 4, and then we have learned a Bayesian network from that sample using the standard K2 algorithm [5]. The result is in Figure 1. We can see a graph with a complex structure and this has been estimated from a very small sample coming from independent variables.

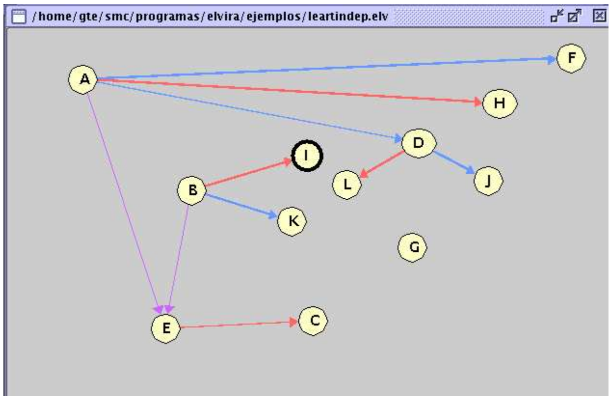

Figure 1: Network learned with a Bayesian score and a sample of size 4 from 12 independent variables.

This problem is not shared by Chi-square independent tests, as for small samples they have a tendency to decide for independence. In fact when using PC learning algorithm [15], which is based on independence tests, we obtain a completely unconnected network from the same sample. But Chi-square tests show another characteristic which is not satisfactory from our point of view: with large samples they keep constant the error of deciding for dependence when there is independence. But, due to the sample size this error could be decreased without a great increase in the other error (assuming independence when there is dependence).

Independence can be studied for different aims: to know about the existence of a relation between two variables or as a previous step to estimate the joint probability distribution (if there is independence this distribution is obtained by estimating the marginal distributions and then taking the product). Also the study of independence could be used as a previous step in classification, for example, if we are building a classification tree [2], we can determine a variable for branching between those that are dependent of the class variable. This paper mainly concentrates in determining the existence of independence, but in the experiments we will give some results relative to the classification problem. There is another related problem which is the measurement of the degree of dependence between two categorical variables [6], to which the imprecise Dirichlet model has been applied [3, 4]. Deciding about independence and measuring the association between two variables are related issues: in most of the cases, the decision is made taken into account a measurement of the dependence degree of the variables, called a dependence score. However, in the paper this score will be always auxiliary to the main aim: determine the existence of independence.

The dependence problem is well known in the Statistical literature [6], and several association measures have been proposed and statistical tests have been designed taking them as basis. However, in this paper we have concentrated in the procedures that have been used in the task of learning Bayesian networks [12] and some new procedures that we have proposed and that are based on the imprecise Dirichlet model.

In this paper we introduce a new imprecise score measure which has its starting point in the Heckerman, Geiger, and Chickering's score [7], but considering an imprecise prior Dirichlet model instead of a precise one [17]. In this way we obtain three possible situations: dependence dominates, independence dominates, or none of them dominates the other. If we decide for independence in all the situations except when dependence dominates, then the behaviour we obtain is intermediate between Chi-square tests (similar to it for small samples) and the Bayesian scores (similar to it for large samples).

In this paper we have carried out a set of experiments to determine the characteristics of the different methods to decide for independence. The setting is very simple: two binary categorical variables and we have to decide whether they are dependent or independent. More complex scenarios are necessary (more than two values, conditional independence with more variables) but, at this stage we only want to discover the most basic behaviour of the procedures and then we have tried to avoid other factors affecting it.

We have compared different classical scores, a score based on imprecise probability proposed in Abellán and Moral [2], a Bayesian score with a prior probability favouring independence, and the imprecise Dirichlet score introduced in this paper.

Though the experiments are with binary variables, our future objective is more important. Our final aim is to determine a general procedure to determine the complexity of a model relating a set of variables, as for example a dependence graph [12] or a classification tree [13]. This construction will be based on the verification of conditional independence relationships between the problem variables, which can have more than two values. So will consider methods that are suitable for being generalized to general categorical variables and to conditional independence relationships. This is a reason for not considering Fisher exact test [6].

The paper is organized as follows: in Section 2, we consider the basic notation and give the basic procedures to decide about independence. In Section 3, we introduce the new interval-valued score. In Section 4, we describe the experiments in detail and the results obtained. The discussion of the results is in Section 5, while Section 6 is devoted to the conclusions and future work.

2 Score Measures

Let X and Y be two variables taking values on set $\{0,1\}$. We assume that there is a probability distribution, p, with which these variables take jointly their values and that p_X and p_Y are its marginal distributions. We also have a sample of pairs of values $(x_1,y_1),\ldots,(x_N,y_N)$ of independent cases, all of which follow distribution p. N is the sample size. The basic problem is to determine from the sample whether there is independence in distribution p. Another important related problem is to obtain an estimation of the joint probability distribution.

To fix the notation, let $n(i,j)$ be the number of occurrences of the pair (i,j) in the sample, and $n_X(i)$

$(n_Y(j))$ the number of times that $X = i$ $(Y = j)$ in the sample. Let us consider that $\hat{p}$ is the maximum likelihood estimation of the distribution ($\hat{p}(i,j)$ is $n(i,j)$ divided by N). Analogously, $\hat{p}_X$ and $\hat{p}_Y$ are the maximum likelihood estimators of the marginal distributions.

All the procedure to decide about dependence are based on a score of the degree of dependence. The first measure of dependence (CHI) that we are going to consider is the p-value of the Chi-square test for independence. This measure is based on considering the estimated mutual information in the sample:

$$G = \sum_{i,j} \hat{p}(i,j) \log \left(\frac{\hat{p}(i,j)}{\hat{p}_X(i)\hat{p}_Y(j)} \right)$$

It is known that, under the independence hypothesis, $2NG$ asymptotically follows a Chi-square distribution with one degree of freedom (as the variables are binary). The p-value is the probability that a Chi-square distribution with 1 degree of freedom is greater than or equal to $2NG$. Our first score of dependence, CHI, will be 1 minus this p-value. It is a number between 0 and 1. Values very close to 1 imply a high level of dependence between the variables. To decide about dependence or independence, we need a significance level α. If $1 - CHI < \alpha$, then we assume that there is independence. The value of α is the probability of the error of deciding dependence when there is independence.

This measure is well justified from classical frequentist statistical theory. It is simple to compute, but it has some drawbacks. The first one is that it is difficult to extend it to a general measure to test a set of independence relationships when we have more than two variables, and therefore there is not an obvious way of obtaining a global score for a general Bayesian network. We could think of considering something as 1 minus the p-value of a Chi-square test in which the null hypothesis is that a set of independence relationships is given, but then this measure will be larger if we have less independence relationships and we will not be penalizing the complexity of the model.

The second measure is based on the K2 score for Bayesian networks [5]. To score independence, it considers the probability distributions p_X and p_Y and then it assumes that there is independence between X and Y and that the parameters of these distributions $(p_X(0), p_X(1))$ and $(p_Y(0), p_Y(1))$ follow two independent Beta distribution with parameters $(1, 1)$ [12]. Then the probability of the sample given these hypotheses is computed being equal to:

$$\begin{aligned} K2I \;=\; & \tfrac{\Gamma(2)}{\Gamma(N+2)} \left(\prod_i \tfrac{\Gamma(n_X(i)+1)}{\Gamma(1)} \right) \\ & \cdot \tfrac{\Gamma(2)}{\Gamma(N+2)} \left(\prod_j \tfrac{\Gamma(n_Y(j)+1)}{\Gamma(1)} \right) \end{aligned}$$

where $\Gamma(x)$ is the gamma function ($\Gamma(n) = (n-1)!$ when applied to an integer as in this case). This value is obtained by integrating the likelihood function with respect to the prior probability for the parameters and it is also called the *marginal likelihood* [12].

To score the dependence case, we consider that the parameters for the marginal distribution $(p_X(0), p_X(1))$ and conditional distributions: $(p_Y(0|X = 0), p_Y(1|X = 0))$ and $(p_Y(0|X = 1), p_Y(1|X = 1))$ follow three independent prior Beta densities with parameters $(1, 1)$, then the probability of the sample can be obtained as

$$\begin{aligned} K2D \;=\; & \tfrac{\Gamma(2)}{\Gamma(N+2)} \left(\prod_i \tfrac{\Gamma(n_X(i)+1)}{\Gamma(1)} \right) \\ & \cdot \prod_i \tfrac{\Gamma(2)}{\Gamma(n_X(i)+2)} \left(\prod_j \tfrac{\Gamma(n(i,j)+1)}{\Gamma(1)} \right) \end{aligned}$$

$K2D$ is also computed by integrating the likelihood with respect to the prior distribution of the parameters.

The final score is $K2 = K2D/K2I$. The decision rule for dependence is that $K2 > 1$.

This score only considers the probability of the data (the sample) given each one of the two possible hypotheses. But, it can be seen as a Bayesian procedure in which we use the posterior probabilities of dependence-independence given the data, under a prior probability of 0.5 for both of them, as in this case $K2I$ and $K2D$ are proportional to the posterior probability of independence and dependence, given the data.

This measure can be used to score general Bayesian networks, as in fact it is a specialization of the K2 measure proposed by Cooper and Herskovits [5]. One of the main criticism to this rule is that its value can depend on the order of the variables. This is due to the fact that if we change the order of the variables the assumptions about the prior distributions of the parameters are not consistent. To avoid this problem, Heckerman, Geiger, and Chickering [7] proposed the so called *Bayesian Dirichlet equivalent scores*. The main question is about the parameters of the Dirichlet prior densities (Beta densities in our case as the variables have two possible values). Simple examples of these measures to score Bayesian networks can be obtained by assuming a global parameters S and then to assume that each conditional density has parameters that are equal to S divided by the product of the number of cases of the variable and the number of configurations of parents variables. In our case with two binary variables, for a value of the parameter S, we obtain the following scores for independence and dependence:

$$\begin{aligned} BSI &= \frac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_i \frac{\Gamma(n_X(i)+S/2)}{\Gamma(S/2)}\right) \\ &\cdot \frac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_j \frac{\Gamma(n_Y(j)+S/2)}{\Gamma(S/2)}\right) \end{aligned}$$

$$\begin{aligned} BSD &= \frac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_i \frac{\Gamma(n_X(i)+S/2)}{\Gamma(S/2)}\right) \\ &\cdot \prod_i \frac{\Gamma(S/2)}{\Gamma(n_X(i)+S/2)} \left(\prod_j \frac{\Gamma(n(i,j)+S/4)}{\Gamma(S/4)}\right) \end{aligned}$$

The value BSD/BSI is the measure of the degree of dependence. If it is greater than 1 we will consider than X and Y are dependent. In our experiments, we will consider three scores $BS0.02, BS2$ and $BS16$ corresponding to values of $S = 0.02, S = 2$ and $S = 16$, respectively.

Another important principle that has been used to determine the complexity of a model is the minimum description length principle [14]. This has been applied to score general probabilistic models [16] giving rise to the so called Bayesian information criterion. It has two components, one measures the fitting of the model to the data and the other penalizes the complexity of the model. In our particular case, we can obtain a measure by considering the difference of the score of the full model and the score of the model with independence. The final expression is:

$$\begin{aligned} BIC &= N \sum_{i,j} \hat{p}(i,j) \log \left(\frac{\hat{p}(i,j)}{\hat{p}_X(i) \cdot \hat{p}_Y(j)}\right) \\ &- (1/2) \log(N) \end{aligned}$$

This model can be also obtained as a Bayesian criterion by selecting some particular prior distribution [10].

In [11] it has been introduced a score based on the theory of imprecise probability [17] and more concretely on the imprecise Dirichlet model. This model is used to compute the probabilities of a categorical probability distribution for Z taking values on set $\{z_1, \ldots, z_k\}$. If we have a sample of size N, and we want to know $p(z_i)$, then it is assumed that the parameters $(p(z_1), \ldots, p(z_k))$ follow an imprecise Dirichlet model with parameter (equivalent sample size) S. With this, we assume as prior information the complete set of Dirichlet distributions with parameters $(\alpha_1, \ldots, \alpha_k)$ where $\alpha_i > 0, \sum_i \alpha_i = S$. The expectation of $p(z_i)$ is a probability interval which is obtained by computing all the posterior expected values of the parameters given all the prior probability distributions. If $n(i)$ is the number of cases for which $Z = z_i$ in the sample, then the interval for $p(z_i)$ will be equal to:

$$\left[\frac{n(i)}{N+S}, \frac{n(i)+S}{N+S}\right]$$

In our case, we will use a value of $S = 1$. Reasons for it are given in [18], but any other positive values is also possible. One important thing is that intervals are wider if the sample size is smaller. So this method produces more precise intervals as N increases.

The entropy of this set of intervals will be measured as the maximum of the entropy of all probability distributions $(q(z_1), \ldots, q(z_k))$ verifying that for any z_i, $q(z_i)$ belongs to the estimated interval for $p(z_i)$. This entropy is simple to compute. First, we have to determine $A = \{z_j \ : \ n(j) = \min_i\{n(i)\}\}$. If l is the number of elements of A, then the distribution with maximum entropy is p^*, where $p^*(z_i) = \frac{n(i)}{N+S}$ if $z_i \notin A$ and $p^*(z_i) = \frac{n(i)+S/l}{N+S}$ if $z_i \in A$. This upper entropy value, denoted as $H^*(Z)$, is an information/uncertainty measure. A justification for its use in general credal sets can be found in [1, 2]. As the intervals are wider with smaller sample sizes, then we will have a tendency to obtain greater values of maximum entropy with smaller sample sizes. It also increases with the number of possible values of variable Z (higher increasing of entropy with respect to the traditional point estimation) especially with very small samples as then $1/(N+1)$ of probability will be uniformly distributed between several cases.

If to determine the values of Z we only consider the part of the sample for which another variable, U takes a value u, then the value of upper entropy will be denoted by $H^*(Z|U = u)$.

The basic intuition of the score is to consider the two cases: independence and dependence. In the case of independence we apply the imprecise probability model to X and Y obtaining the entropy of the global model as $IPI = H^*(X) + H^*(Y)$. In the case of dependence of Y from X, we apply the imprecise probability model to X and to each one of the conditional probabilities about Y, given $X = 0$ and given $X = 1$. In this stage we could compute the upper entropy of all the distributions that can be obtained by composing[1] the sets of distributions obtained in this way. However, this poses a non simple computational problem (at least for the general case of more than two variables) and we have considered as upper entropy of the dependence case the value:

[1] The composition is the usual multiplication of marginal and conditional probability distributions.

$$IPD = H^*(X) + \sum_i \hat{p}(i).H^*(Y|X=i)$$

Finally the imprecise probability score is the upper entropy of independence minus the upper entropy of dependence: $IMP = IPI - IDP = H^*(Y) - \sum_i \hat{p}(i).H^*(Y|X=i)$. We decide for dependence if $IMP > 0$ and for independence in other case. This score, resembles the mutual information degree of dependence, but it can be lower than 0. The basic idea is to measure the uncertainty under dependence and independence and then to chose the situation with lower uncertainty.

This measure can be extended to score general Bayesian networks. However, it is not symmetrical in the variables. The final value of IMP depends of the order of the variables. This is due to the fact that assuming a global imprecise Dirichlet model for X and for all the conditional distributions of Y given X is not equivalent to assume it for Y and for the conditional distributions of X given Y.

3 The New Imprecise Score Measure

The new score is based on Bayesian equivalent score and the imprecise Dirichlet model. Instead of equidistributing the S value among all the possible elements, it considers a family of parameter values for a S value and test whether independence dominates dependence for all of them, or vice versa. For the sake of simplicity, we introduce it for the particular case of two binary variables, but its generalization to more cases per variable is immediate.

We assume that we have a fixed S value. Then we consider that the prior information about $p(i,j)$ is a Dirichlet distribution of parameters $\alpha = (\alpha(0,0), \alpha(0,1), \alpha(1,0), \alpha(1,1))$, where $\sum_{i,j} \alpha(i,j) = S$. The probability of the data (marginal likelihood), under this prior information is given by:

$$\begin{aligned} BSD_\alpha \quad = \quad & \tfrac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_i \tfrac{\Gamma(n_X(i)+\alpha_X(i))}{\Gamma(\alpha_X(i))}\right) \\ & \cdot \prod_i \tfrac{\Gamma(\alpha_X(i))}{\Gamma(n_X(i)+\alpha_X(i))} \left(\prod_j \tfrac{\Gamma(n(i,j)+\alpha(i,j))}{\Gamma(\alpha(i,j))}\right) \end{aligned}$$

Where $\alpha_X(i) = \sum_j \alpha(i,j), \quad \alpha_Y(j) = \sum_i \alpha(i,j)$.

Under independence, we assume that $p(0|i) = p(1|i)$ in the prior information and then the probability of the data under the resulting distribution is given by

$$\begin{aligned} BSI_\alpha \quad = \quad & \tfrac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_i \tfrac{\Gamma(n_X(i)+\alpha_X(i))}{\Gamma(\alpha_X(i))}\right) \\ & \cdot \tfrac{\Gamma(S)}{\Gamma(N+S)} \left(\prod_j \tfrac{\Gamma(n_Y(j)+\alpha_Y(j))}{\Gamma(\alpha_Y(j))}\right) \end{aligned}$$

Considering a set $\Phi \subseteq \{\alpha \mid \alpha(i,j) > 0, \sum_{i,j} \alpha(i,j) = S\}$, the new interval-valued score can be defined as:

$$[BS_*, BS^*] = \left[\min_{\alpha \in \Phi} \frac{BSD_\alpha}{BSI_\alpha}, \max_{\alpha \in \Phi} \frac{BSD_\alpha}{BSI_\alpha}\right]$$

We assume that dependence dominates if $BS_* > 1$, independence dominates when $BS^* < 1$ and there is no dominance when $1 \in [BS_*, BS^*]$. This agrees with the usual dominance for imprecise probability [17] under strict preference taking as basis the posterior probability of having dependence or independence (considering that the prior probability is 0.5 for both of them). This criterion will be denoted as $BSDOM$.

For set Φ we have not considered the full set of possibilities: $\{\alpha \mid \alpha(i,j) > 0, \sum_{i,j} \alpha(i,j) = S\}$ as in the imprecise Dirichlet model considered in [18]. The reason is that, with the exception of some trivial cases, we can make BSD_α / BSI_α as small as we want (by taking a very small $\alpha(i,j)$, whereas $\alpha_X(i)$ and $\alpha_Y(j)$ are not so small, when at least in one observation we have $X = i, Y = j$). Then the lower limit of the interval would always approach 0 and dependence would never dominate. We have found that a reasonable approach is to divide the S value in two parts $S = S_1 + S_2$. Then S_1 is uniformly split and we consider all the possible parameters for S_2 value. Being more specific, if $\alpha_1(i,j) = S_1/4$ and $\Phi_2 = \{\alpha_2 \mid \alpha_2(i,j) > 0, \sum_{i,j} \alpha_2(i,j) = S_2\}$, then $\Phi = \{\alpha = \alpha_1 + \alpha_2 \mid \alpha_2 \in \Phi_2\}$ where the addition of vectors is pointwise addition. All the experiments in this paper corresponds to $S_1 = S_2 = S/2$ and $S = 2$.

We do not know any direct method to compute the upper and lower extremes of the interval, BS_* and BS^*. It is also possible that if we consider the convex hull of the family of prior distributions, then the upper and lower values of the interval will change (we are optimizing a function for which we have not security that the optimum is obtained in an extreme point). However, we will keep only the Dirichlet prior distributions (we consider a non convex prior information). In this set, we have carried out an approximate computation[2]. To compute the lower extreme, we have selected the parameters α_2 trying to favour independence as much as possible. We have considered the following two possible selections for α_2 (α is computed as $\alpha_1 + \alpha_2$):

1. For each value $i = 0, 1$, if $n(i,j) > n(i,j')$, then assign $\alpha_2(i,j') = S_2/2, \alpha_2(i,j) = 0$; if $n(i,j) = n(i,j')$, then $\alpha_2(i,j') = \alpha_2(i,j') = S_2/4$.

[2] We do not believe that it is really a very hard problem, and in most of the cases the extremes of the intervals are obtained with extreme values of parameters α_2.

2. For each value $i = 0, 1$, if $n(i,j)/n_X(i) > n_Y(j)/N$, then assign $\alpha_2(i,j') = S_2/2, \alpha_2(i,j) = 0$; if $n(i,j)/n_X(i) = n_Y(j)/N$, then $\alpha_2(i,j') = \alpha_2(i,j') = S_2/4$.

The first strategy tries to make the conditional distributions of Y given X as uniform as possible. The second tries to makes the conditional distributions as similar as possible to the marginal one. They are similar, but they do not always assign parameters in the same way. For both cases, we compute the values BSD_α/BSI_α, and take the minimum of them.

To compute the upper interval limit, we also consider two parameters but trying to favour dependence. We have also considered two vectors of parameters, computing the maximum of BSD_α/BSI_α for both vectors of them:

1. For each value $i = 0, 1$, if $n(i,j) \geq n(i,j')$, then assign $\alpha_2(i,j) = S_2/2, \alpha_2(i,j') = 0$.

2. For each $i = 0, 1$, if $n(i,j)/n_X(i) \geq n_Y(j)/N$, then assign $\alpha_2(i,j) = S_2/2, \alpha_2(i,j') = 0$.

The strategy is dual of the strategy for the lower limit.

One of the criteria we are going to test in the experiments, is to consider independence if $BS_* \geq 1$ and dependence otherwise. This criterion always makes a decision and follows the intuitive idea of selecting independence except when we have evidence enough favouring dependence. It is similar to frequentist tests of hypothesis, where the null hypothesis is accepted except if we have evidence against it. In our case, if the data do not decide for independence or dependence, then we should prefer the simpler model that does not assume the existence of a relation between the two variables. However, as we will see it will have better asymptotic properties than classical statistical tests.

4 Experiments

We have carried out two series of experiments. In both of them, we have simulated 10000 joint probability distributions for (X, Y) with dependence and 10000 in which X and Y are independent. To obtain the probabilities we have followed Dirichlet distributions. In the two cases, the procedure has been the same with the only difference of S value. We have considered the case of $S = 2$ and $S = 8$. In the dependence case the probabilities are randomly simulated according to a Dirichlet distribution of parameters $(S/4, S/4, S/4, S/4)$. In the independence case, we follow a similar procedure to obtain the marginal distributions (simulated according to a Beta of parameters $(S/2, S/2)$) and then the joint probability is computed as product of the marginal distributions.

For each one of the distributions, we have simulated samples with sizes: 3, 5, 10, 20, 50, 100, 1000, 10000. We consider very small samples as this is a particularly important case. Even if we have a large sample, when we are going to determine a model for the data, and its complexity is going to depend of the amount of data as in Bayesian networks learning, then some crucial decisions are generally done with small samples.

Then we have tried to determine from each sample the existence of dependence or independence of variables X and Y. For each one of the scores we have measured the number of errors in recovering dependence-independence relationships (for the dominance criterion, $BSDOM$, we also give the number of cases in which there is a decision). But, to evaluate their use in classification, we have also computed the average of the expected log-likelihood of the estimated probability of Y given X with respect to the true probability distribution, according to the following expressions:

1. *If deciding for dependence* $\sum_{i,j} p(i,j) \log p^*(j|i)$

2. *If deciding for independence* $\sum_{i,j} p(i,j) \log p^*(j)$

where p is the original distribution with which the samples were simulated and $p^*(i,j) = \frac{n(i,j)+0.5}{N+2}$, $p^*(j) = \frac{n_Y(j)+1}{N+2}$.

As larger is this value, the better is the method for deciding whether X is useful to estimate the probability of Y.

We report first and with more detail the results for $S = 2$. In this case, we follow exactly the same hypotheses to generate the distributions that are assumed by $BS2$ score, so we are in the ideal situation for this score, and it should outperform the other ones. The tables contain the results for all the methods to decide about independence we have introduced and another score $BS2MOD$ which is equal to $BS2$, but modifying the prior distribution for independence - dependence. In $BS2$ it was 1/2 for each one of the two possibilities. In $BS2MOD$ we have given more prior information to independence, to the point of obtaining exactly the same behaviour of BS_* for small samples (3 and 5). The number of errors can be seen in Table 1 for the case of true dependence and in Table 2 in the case of distributions generated with independence of X and Y. $BSDOM$ criterion includes the number of decisions between parentheses.

Tables 3 and 4 show the average expected log-

Table 1: Number of errors for 10000 repetitions for each sample size (independence and $S = 2$).

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM (Dec.)
3	0	839	839	3346	3346	3346	0	839	0	0 (6654)
5	312	1553	1553	2641	2641	4832	86	1553	86	86 (7445)
10	392	1287	1991	1258	2043	3691	110	2099	529	529 (8174)
20	506	792	1631	751	1546	3878	134	2356	540	540 (8770)
50	492	466	1215	420	1029	3439	120	2595	463	463 (9233)
100	523	313	896	279	726	2915	84	2636	397	397 (9486)
1000	507	84	281	55	223	1195	34	2686	145	145 (9920)
10000	482	27	76	13	71	347	12	2628	53	53 (9987)

Table 2: Number of errors on 10000 repetitions for each sample size (dependence and $S = 2$).

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM (Dec.)
3	10000	8142	8142	5078	5078	5078	10000	8142	10000	5078 (5078)
5	8290	6750	6750	5141	5141	3287	9211	6750	9211	5141 (5930)
10	7225	5948	5456	5654	4822	3651	7985	5802	7050	4548 (7498)
20	5783	5359	4635	5748	4259	2844	6760	4585	5846	4209 (8363)
50	4136	4177	3568	5012	3382	2263	5123	3315	4302	3232 (8930)
100	3130	3400	2868	4259	2786	1872	3980	2522	3292	2704 (9412)
1000	1112	1453	1277	1898	1206	875	1591	983	1290	1204 (9914)
10000	383	557	513	711	489	393	580	366	509	493 (9984)

likelihood for the cases of dependence and independence, respectively (in the case of dominance this average is only for the cases in which there is a decision).

From Tables 1, 2 and Tables 3, 4, we have obtained Tables 5 and 6 respectively, where we give the addition of the errors and the average log-likelihood for independence and dependence together.

The results for $S = 8$ are given in tables 7 and 8, but integrating the cases of dependent and independent distributions.

5 Discussion

We summarize our analysis of experiments results in the following points:

- First we highlight an important property of the procedure based on BS_*. With small sample sizes it always decides for independence. This is similar to the chi-square tests, where if there is not evidence against independence, then it decides for independence. However, when the sample size increases, then it is more similar to the Bayesian score. Test of independence keeps the error of deciding dependence when there is independence constant (around 5% in our case). But, when sample size increases this error can be reduced without a big effect into the dual error. This is achieved by the BS_* score that it is very similar to the $BS2$ score for large sample sizes. The main difference with CHI is that BS_* reduces the number of errors of assuming dependence when there is independence for large sample sizes.
- There is no procedure which is better than the others ones in all the situations, though Bayesian score $BS2$ shows a very good behaviour in our experiments in relation with the number of errors. This is not surprising as we generated the distributions of the experiments with exactly the hypotheses assumed by $BS2$ in one case. In the other case, we also used Dirichlet distributions to generate the distributions, but with a different S value.
- In general, a smaller number of errors implies a bigger log-likelihood, but this is not always the case. The main reason is that the log-likelihood errors are not symmetrical from dependence to independence and vice versa. So the final average error will depend of whether a method favours dependence or independence. Also the scores have been designed in order to detect independence relationships and not for classification purposes (except perhaps IMP that follows a more classification oriented criterion). For classification, Bayesian procedures should be reformulated in order to take into account the different classification errors. Observe as score BS_* obtains better log-likelihood values for the experiments in which $S = 8$ especially for small and intermediate samples. BS_* is also based on a value of $S = 2$ as $BS2$ but obtains better results by taking more

Table 3: Average log-likelihood for 10000 repetitions for each sample size (independence and $S = 2$).

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM
3	-0.587167	-0.603542	-0.603542	-0.617858	-0.617858	-0.617858	-0.587167	-0.603542	-0.587167	-0.607798
5	-0.571321	-0.583437	-0.583437	-0.589336	-0.589336	-0.592899	-0.565971	-0.583437	-0.565971	-0.572510
10	-0.544609	-0.551856	-0.555106	-0.550760	-0.553739	-0.557997	-0.540602	-0.552790	-0.543422	-0.545230
20	-0.526979	-0.528688	-0.531910	-0.527086	-0.531123	-0.533776	-0.523701	-0.531798	-0.526012	-0.525874
50	-0.511454	-0.511377	-0.513114	-0.510493	-0.512561	-0.514426	-0.509869	-0.514152	-0.511020	-0.515999
100	-0.505875	-0.505473	-0.506311	-0.505029	-0.506031	-0.507035	-0.504841	-0.507158	-0.505492	-0.510250
1000	-0.500295	-0.500191	-0.500239	-0.500165	-0.500224	-0.500321	-0.500171	-0.500451	-0.500204	-0.501925
10000	-0.499720	-0.499708	-0.499710	-0.499707	-0.499710	-0.499715	-0.499707	-0.499736	-0.499709	-0.500224

Table 4: Average log-likelihood for 10000 repetitions for each sample size (dependence and $S = 2$).

Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM
3	-0.586945	-0.553335	-0.553335	-0.533924	-0.533924	-0.533924	-0.586945	-0.553335	-0.586945	-0.608972
5	-0.523290	-0.509295	-0.509295	-0.498596	-0.498596	-0.495148	-0.542460	-0.509295	-0.542460	-0.540739
10	-0.475576	-0.465054	-0.462433	-0.463466	-0.460152	-0.456599	-0.485666	-0.466862	-0.478404	-0.450949
20	-0.434797	-0.432386	-0.429274	-0.439151	-0.427918	-0.426349	-0.442979	-0.430384	-0.436730	-0.417088
50	-0.406696	-0.406789	-0.405274	-0.413848	-0.405004	-0.404112	-0.410667	-0.405381	-0.407574	-0.397799
100	-0.396217	-0.396703	-0.395917	-0.400719	-0.395829	-0.395353	-0.398164	-0.395814	-0.396629	-0.393210
1000	-0.386563	-0.386643	-0.386604	-0.386897	-0.386590	-0.386554	-0.386693	-0.386567	-0.386608	-0.386998
10000	-0.385631	-0.385636	-0.385635	-0.385646	-0.385634	-0.385632	-0.385638	-0.385632	-0.385635	-0.385888

possibilities for the α parameters into account.

- In order to test whether BS_* behaviour can be obtained with a pure Bayesian procedure, we have considered $BS2MOD$ in which prior probability for independence has been increased in order to mimic BS_* for that sizes. However, the asymptotic behaviour of BS_* is better: the total number of errors is always lower with BS_* when the sample size increases.

- Procedures IMP and CHI makes too many errors for large sample sizes, especially by no decreasing the errors of deciding dependence when there is independence, but these errors are not so important when we look at the log-likelihood. The reason is that with a large sample size, we can get good estimations of the joint probability distribution without assuming the existing independence. Independence is more important for small sample sizes, but then the number of errors is similar to the Bayesian procedures. So these methods are appropriate for classification problems. Furthermore, they do not make additional hypotheses about the way in which data are generated and their behaviour is not expected to deteriorate when these hypotheses are not fulfilled.

- Dominance criterion makes a lower percentage of errors that methods that make decisions in all the cases, but this procedure avoids to classify unsure situations. When looking at the average log-likelihood we can observe that, in general, it is also decreased, but there is an apparent paradoxical result for sample size 5 and $S = 2$. We obtained a lower average log-likelihood by using the dominance criterion. This surprised us, as classifying only the sure cases we expected to increase the average log-likelihood. But, at the end we understand that this is a plausible result. The explanation is that dominance only takes into account independence-dependence errors which are measured in a symmetrical way. If we want to maximize log-likelihood all the problem should be reformulated for this aim, and then the decision procedures would be different. It is possible that even if we do fewer errors in percentage the errors we are doing have a relatively low log-likelihood in relation with the cases in which we do not make decisions. In any case, there is a strong asymmetry in the number of errors of the dominance criterion. Most of the errors are done by considering independence when there is dependence. We do not have a reason for it, and possibly is due to the imprecise Dirichlet model we have considered or to the approximate computation.

- If we compare Bayesian procedures (including K2 and BIC) with BS_* we see that even with these few experiments none of them is better than the others in all the situations. For example, BS_* is better with large samples for $S = 2$ (for log-likelihood and number of errors) and in small samples for $S = 8$ (log-likelihood criterion).

- With larger S values in the Bayesian equivalent criteria, we have more tendency to favour dependence. This can be seen looking at the errors of $BS0.02, BS2$, and $BS16$ in the cases of dependence and independence. However, these differences are less important for small sample sizes. So we would have not obtained a similar score

Table 5: Integrated number of errors for $S = 2$ obtained from Table 1 and Table 2.

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM (Dec.)
3	10000	8981	8981	8424	8424	8424	10000	8981	10000	5078 (11732)
5	8602	8303	8303	7782	7782	8119	9297	8303	9297	5227 (13375)
10	7617	7235	7447	6912	6865	7342	8095	7901	7579	5077 (15672)
20	6289	6151	6266	6499	5805	6722	6894	6941	6386	4749 (17133)
50	4628	4643	4783	5432	4411	5702	5243	5910	4765	3695 (18163)
100	3653	3713	3764	4538	3512	4787	4064	5158	3689	3101 (18898)
1000	1619	1537	1558	1953	1429	2070	1625	3669	1435	1349 (19834)
10000	865	584	589	724	560	740	592	2994	562	546 (19967)

Table 6: Integrated averaged log-likelihood for $S = 2$ obtained from Table 3 and Table 4.

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM
3	-0.587056	-0.5784385	-0.5784385	-0.575891	-0.575891	-0.575891	-0.587056	-0.5784385	-0.587056	-0.608306
5	-0.5473055	-0.546366	-0.546366	-0.543966	-0.543966	-0.5440235	-0.5542155	-0.546366	-0.5542155	-0.558423
10	-0.5100925	-0.508455	-0.5087695	-0.507113	-0.5069455	-0.507298	-0.513134	-0.509826	-0.510913	-0.500123
20	-0.480888	-0.480537	-0.480592	-0.4831185	-0.4795205	-0.4800625	-0.48334	-0.481091	-0.481371	-0.472773
50	-0.459075	-0.459083	-0.459194	-0.4621705	-0.4587825	-0.459269	-0.460268	-0.4597665	-0.459297	-0.457885
100	-0.451046	-0.451088	-0.451114	-0.452874	-0.45093	-0.451194	-0.4515025	-0.451486	-0.4510605	-0.451959
1000	-0.443429	-0.443417	-0.4434215	-0.443531	-0.443407	-0.4434375	-0.443432	-0.443509	-0.443406	-0.444479
10000	0.4426755	-0.442672	-0.4426725	-0.4426765	-0.442672	-0.4426735	-0.4426725	-0.442684	-0.442672	-0.443053

to BS_* by considering dominance under several values of parameter S. In fact, the lower limit of the dominance interval would be equal or very similar to the Bayesian score with the lowest S value ($BS0.02$) which makes a lot of errors, by considering dependence in the case of independence, with a sample size of 3.

6 Conclusions and Future Work

In this paper we have carried out an empirical comparison of several criteria for deciding independence and introduced a new method based on the theory of imprecise probability. Perhaps the most important conclusion is that no single method outperforms the others. The final criterion should be chosen as a function of the objective (deciding independence or classification) and the sample size.

But, the main conclusion of this paper is a new procedure to decide for dependence-independence that for small sample sizes always considers independence and that for large sample sizes is similar to Bayesian procedures. This method can be used when we want to determine the dependence relationships that can be found in a set of data, but we really only want relationships with real support, avoiding spurious relationships.

For future work we plan to consider the following points:

- To make more extensive studies, changing the procedure of generating the distributions, and studying the errors in the case of dependence as a function of the mutual information between the variables (measured in the original true distribution).
- To study the performance of the criteria for variables with more than two possible values.
- To consider new criteria or modifications of the existing ones. For example, to study different sets of parameters for the imprecise Dirichlet model.
- To consider criteria based on the exact distribution of the mutual information given the data taking into account the results by Hutter and Zaffalon [8].
- To extend BS_* criterion to be used in learning Bayesian networks algorithms.
- To determine classification oriented dominance criteria based on imprecise probability.

References

[1] J. Abellán and S. Moral (2003) Maximum entropy for credal sets, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11, 587–597.

[2] J. Abellán, S. Moral (2005) Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning* 39, 235–255.

Table 7: Errors for each sample size (10000 repetitions with independence and 10000 with dependence, $S = 8$).

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM (Dec.)
3	10000	9397	9397	9290	9290	9290	10000	9397	10000	6671 (14052)
5	9356	8996	8996	8857	8857	8868	9626	8996	9626	6280 (14078)
10	8674	8496	8348	8583	8386	8202	9116	8823	9015	6441 (15459)
20	7751	7451	7261	8685	7371	7265	8356	8048	8077	6661 (17659)
50	6100	6118	5897	8150	6013	5907	6981	7087	6451	5876 (19161)
100	4750	4903	4671	6946	4787	4679	5732	6407	5123	4861 (19636)
1000	1952	2038	1939	2859	2015	1940	2313	5025	2085	2066 (19975)
10000	985	783	746	1035	773	774	867	4604	783	781 (19998)

Table 8: Averaged log-likelihood for each sample size (10000 repetitions with independence and 10000 with dependence, $S = 8$).

Sample Size	CHI	BIC	K2	BS0.02	BS2	BS16	BS2MOD	IMP	BS$_*$	BSDOM
3	-0.7052005	-0.7223925	-0.7223925	-0.7236235	-0.7236235	-0.7236235	-0.7052005	-0.7223925	-0.7052005	-0.703886
5	-0.698873	-0.7124025	-0.7124025	-0.7120575	-0.7120575	-0.711724	-0.696166	-0.7124025	-0.696166	-0.698185
10	-0.6761785	-0.6814055	-0.683148	-0.6791815	-0.679737	-0.681291	-0.671937	-0.6812355	-0.672429	-0.671644
20	-0.6493725	-0.649894	-0.650813	-0.652661	-0.650037	-0.650153	-0.6483265	-0.6514565	-0.648816	-0.648923
50	-0.6261545	-0.6261475	-0.6261755	-0.6311765	-0.626154	-0.6260925	-0.6274595	-0.626938	-0.6265345	-0.625931
100	-0.616961	-0.6170425	-0.6169415	-0.620509	-0.6169615	-0.6169375	-0.618013	-0.617455	-0.6172415	-0.616807
1000	-0.6086205	-0.608631	-0.6086175	-0.6088435	-0.6086265	-0.608617	-0.6086795	-0.6087205	-0.608638	-0.608612
10000	-0.60784	-0.607838	-0.6078375	-0.607846	-0.607838	-0.6078375	-0.6078405	-0.6078525	-0.607838	-0.607831

[3] J.M. Bernard (2003) Analysis of Local or Asymmetric Dependencies in Contingency Tables using the Imprecise Dirichlet Model. In: *Proc. of ISIPTA'03* (J. Bernard, T. Seidenfeld, M. Zaffalon, eds.) Lugano (Suiza) 46–61.

[4] J.M. Bernard (2005) An Introduction to the Imprecise Dirichlet Model for Multinomial Data. *International Journal of Approximate Reasoning* 39, 123–150.

[5] G.F. Cooper, E. Herskovits (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9, 309–348.

[6] L.A. Goodman, W.H. Kruskal (1980) *Measures of Association for Cross- Classification, Volume 1.* New York, Springer-Verlag.

[7] D. Heckerman, D. Geiger, D.M. Chickering (1995) Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20, 197–243.

[8] M. Hutter, M. Zaffalon (2004) Distribution of Mutual Information from Complete and Incomplete Data. To appear in: *Computational Statistics and Data Analysis.*

[9] S. Kullback (1968) *Information Theory and Statistics.* Dover, Gloucester.

[10] W. Lam, F. Bacchus (1994) Learning Bayesian Belief Networks. An Approach Based on the MDL Principle. *Computational Intelligence* 10, 269–293.

[11] S. Moral (2004) An Empirical Comparison of Score Measure for Independence. In: *Proc. of IMPU'04* vol 2, 1307–1314.

[12] R.E. Neapolitan (2004) *Learning Bayesian Networks.* Upper Saddle River, Prentice Hall.

[13] J.R. Quinlan (1986) Induction of Decision Trees. *Machine Learning* 1, 81–106.

[14] J. Rinassen (1978) Modeling by Shortest Data Description. *Automatica* 14, 465–471.

[15] P. Spirtes, C. Glymour, R. Scheines (1993) *Causation, Prediction, and Search.* Springer-Verlag, New York.

[16] G. Schwarz (1978) Estimating the Dimension of a Model. *Annals of Statistics* 6, 461–464.

[17] P. Walley (1991) *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

[18] P. Walley (1996) Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society*, Series B 58, 3-57.