Entropy **2015**, *17*, 5752-5770; doi:10.3390/e17085752

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Consistency of Learning Bayesian Network Structures with Continuous Variables: An Information Theoretic Approach

Joe Suzuki

Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka-shi 560-0043, Japan; E-Mail: suzuki@math.sci.osaka-u.ac.jp

Academic Editor: Carlo Cafaro

Received: 30 April 2015 / Accepted: 5 August 2015 / Published: 10 August 2015

Abstract: We consider the problem of learning a Bayesian network structure given n examples and the prior probability based on maximizing the posterior probability. We propose an algorithm that runs in $O(n \log n)$ time and that addresses continuous variables and discrete variables without assuming any class of distribution. We prove that the decision is strongly consistent, *i.e.*, correct with probability one as $n \to \infty$. To date, consistency has only been obtained for discrete variables for this class of problem, and many authors have attempted to prove consistency when continuous variables are present. Furthermore, we prove that the "$\log n$" term that appears in the penalty term of the description length can be replaced by $2(1+\epsilon) \log \log n$ to obtain strong consistency, where $\epsilon > 0$ is arbitrary, which implies that the Hannan–Quinn proposition holds.

Keywords: posterior probability; consistency; minimum description length; universality; discrete and continuous variables; Bayesian network

---

1. Introduction

In this paper, we address the problem of learning a Bayesian network structure from examples.

For sets A, B, C of random variables, we say that A and B are conditionally independent given C if the conditional probability of A and B given C is the product of the conditional probabilities of A given C and B given C. A Bayesian network (BN) is a graphical model that expresses conditional independence (CI) relations among the prepared variables using a directed acyclic graph (DAG). We define a BN by the DAG with vertexes $V = \{1, \cdots, N\}$ and directed edges $E = \{(j, i) | i \in V, j \in \pi(i)\}$,

where edge $(j,k) \in V^2$ directs from j to k, via minimal parent sets $\pi(i) \subseteq V$, $i \in V$, such that the distribution is factorized by:

$$P(X^{(1)}, \cdots, X^{(N)}) = \prod_{i=1}^{N} P(X^{(i)} | \{X^{(j)}\}_{j \in \pi(i)}) \, .$$

First, suppose that we wish to know whether two random binary variables X and Y are independent (hereafter, we write $X \perp\!\!\!\perp Y$). If we have n pairs of actually emitted examples $(X = x_1, Y = y_1), \cdots, (X = x_n, Y = y_n)$ and know the prior probability p of $X \perp\!\!\!\perp Y$, then it would be reasonable to maximize the posterior probability of $X \perp\!\!\!\perp Y$ given $x^n = (x_1, \cdots, x_n)$ and $y^n = (y_1, \cdots, y_n)$. If we assume that the probabilities $P(X = x), P(Y = y)$ and $P(X = x, Y = y)$ are parameterized by $p(x|\theta_X), p(y|\theta_Y)$, and $p(x, y|\theta_{XY})$ and that the prior probabilities W_X, W_Y, and W_{XY} over the probabilities θ_X, θ_Y, and θ_{XY} of $X \in \{0,1\}$, $Y \in \{0,1\}$ and $(X,Y) \in \{0,1\}^2$ are available, respectively, then we can construct the quantities:

$$\begin{aligned} Q_X^n(x^n) &:= \int \prod_{i=1}^{n} p(x_i|\theta_X) W_X(d\theta_X) \, , \\ Q_Y^n(y^n) &:= \int \prod_{i=1}^{n} p(y_i|\theta_Y) W_Y(d\theta_Y) \, , \\ Q_{XY}^n(x^n, y^n) &:= \int \prod_{i=1}^{n} p(x_i, y_i|\theta_{XY}) W_{XY}(d\theta_{XY}) \, . \end{aligned}$$

In this setting, maximizing the posterior probability of $X \perp\!\!\!\perp Y$ given examples x^n, y^n w.r.t. the prior probability p is equivalent to deciding $X \perp\!\!\!\perp Y$ if and only if:

$$p Q_X^n(x^n) Q_Y^n(y^n) \geq (1-p) Q_{XY}^n(x^n, y^n) \, . \tag{1}$$

The decision based on (1) is strongly consistent, *i.e.*, it is correct with probability one as $n \to \infty$ [1] (see Section 3.1 for the proof). We say that a model selection procedure satisfies weak consistency if the probability of choosing the correct model goes to unity as n grows (probability convergence) and that it satisfies strong consistency if the probability one is assigned to the set of infinite example sequences that choose the correct model, except for at most finite times (almost sure convergence). In general, strong consistency implies weak consistency, but the converse is not true [2]. In any model selection, in particular for large n, the correct answer is required. If continuous variables are present, the BN structure learning is not easy, and strong consistency is hard to obtain.

The same scenario is applied to the case in which X and Y take values from finite sets A and B rather than $\{0,1\}$.

Next, suppose that we wish to know the factorization of three random binary variables X, Y, Z: $P(X)P(Y)P(Z)$, $P(X)P(Y,Z)$, $P(Y)P(Z,X)$, $P(Z)P(X,Y)$, $\dfrac{P(X,Y)P(X,Z)}{P(X)}$, $\dfrac{P(X,Y)P(Y,Z)}{P(Y)}$, $\dfrac{P(X,Z)P(Y,Z)}{P(Z)}$, $\dfrac{P(Y)P(Z)P(X,Y,Z)}{P(Y,Z)}$, $\dfrac{P(Z)P(X)P(X,Y,Z)}{P(Z,X)}$, $\dfrac{P(X)P(Y)P(X,Y,Z)}{P(X,Y)}$ and $P(X,Y,Z)$. If we have n triples of actually emitted examples $(X = x_1, Y = y_1, Z = z_1), \cdots,$

$(X = x_n, Y = y_n, Z = z_n)$ and know the prior probabilities $p_1, \cdots, p_{11}$ over the eleven factorizations, then it would be reasonable to choose the one that maximizes:

$$p_1 Q_X^n(x^n) Q_Y^n(y^n) Q_Z(z^n), \quad p_2 Q_X^n(x^n) Q_{YZ}^n(y^n, z^n), \quad p_3 Q_Y^n(y^n) Q_{XZ}^n(x^n, z^n),$$
$$p_4 Q_Z^n(z^n) Q_{XY}^n(x^n, y^n), \quad p_5 \frac{Q_{XY}^n(x^n,y^n) Q_{XZ}^n(x^n,z^n)}{Q_X^n(x^n)}, \quad p_6 \frac{Q_{XY}^n(x^n,y^n) Q_{YZ}^n(y^n,z^n)}{Q_Y^n(y^n)},$$
$$p_7 \frac{Q_{XZ}^n(x^n,z^n) Q_{YZ}^n(y^n,z^n)}{Q_Z^n(z^n)}, \quad p_8 \frac{Q_Y^n(y^n) Q_Z^n(z^n) Q_{XYZ}^n(x^n,y^n,z^n)}{Q_{YZ}^n(y^n,z^n)}, \quad p_9 \frac{Q_Z^n(z^n) Q_X^n(x^n) Q_{XYZ}^n(x^n,y^n,z^n)}{Q_{XZ}^n(x^n,z^n)},$$
$$p_{10} \frac{Q_X^n(x^n) Q_Y^n(y^n) Q_{XYZ}^n(x^n,y^n,z^n)}{Q_{XY}^n(x^n,y^n)}, \quad p_{11} Q_{XYZ}^n(x^n, y^n, z^n) \ ,$$

to maximize the posterior probability of the factorization given $x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$ and $z^n = (z_1, \cdots, z_n)$. For example, between the last two distributions, we choose the last if and only if:

$$p_{10} Q_X^n(x^n) Q_Y^n(y^n) \leq p_{11} Q_{XY}^n(x^n, y^n) \ .$$

In fact, for example, we can check that the factorizations:

$$P(Y)P(X|Y)P(Z|X) \ , \ P(X)P(Y|X)P(Z|X) \ , \ P(Z)P(X|Z)P(Y|Z)$$

in Figure 1a–c share the same form $\dfrac{P(XY)P(XZ)}{P(X)}$, and we say that they share the same Markov-equivalent class. On the other hand, the factorization

$$P(Y)P(Z)P(X|YZ) = \frac{P(XYZ)P(Y)P(Z)}{P(YZ)}$$

in Figure 1d has nothing to share with the same Markov equivalent class, except itself. In the case of three variables, there are 25 DAGs, but they reduce to the eleven Markov equivalent classes.

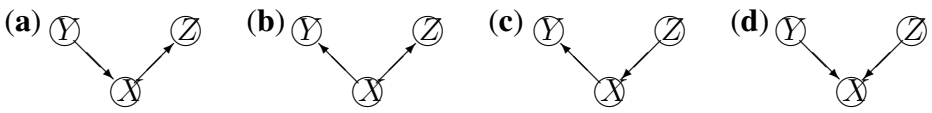

Figure 1. Markov-equivalent classes (**a**–**d**).

The method that maximizes the posterior probability is strongly consistent [1] (see Section 3.1 for the proof), and a scenario with two and three variables as above can be extended to cases with N variables in a straightforward manner, if the variables are discrete.

In this paper, we consider the case when continuous variables are present. The idea is to construct measures $g_X^n(x^n)$, $g_Y^n(y^n)$ and $g_{XY}^n(x^n, y^n)$ over $\mathcal{X}^n$, $\mathcal{Y}^n$ and $\mathcal{X}^n \times \mathcal{Y}^n$ for continuous ranges $\mathcal{X}$ and $\mathcal{Y}$ to make the decision whether $X \perp\!\!\!\perp Y$ based on:

$$p g_X^n(x^n) g_Y^n(y^n) \geq (1-p) g_{XY}^n(x^n, y^n) \ . \tag{2}$$

The main problem is whether the decision is strongly consistent. Many authors have attempted to address continuous variables. For example, Nir Friedman [3] experimentally demonstrated the construction of a genetic network based on expression data using the E-Malgorithm. However, the variables were assumed to be linearly related and included Gaussian noise, and the dataset was not sufficiently fit to the model.

Imoto *et al*. [4] improved the model such that the relation is expressed by B-spline curves rather than lines. However, all of the authors, including Friedman and Imoto, failed to maximize the posterior probability, and thus, the decision is not consistent. This paper proves that the decision based on (2) and its extension for general $N \geq 2$ is strongly consistent.

In any Bayesian approach of BN structure learning, whether continuous variables are present or not, the procedure consists of two stages:

(1) Compute the local scores for the nonempty subsets of $\{X^{(1)}, \cdots, X^{(N)}\}$; for example, if $N = 3$, the seven quantities $Q_X^n(x^n), \cdots, Q_{XYZ}^n(x^n, y^n, z^n)$ are obtained; and
(2) Find a BN structure that maximizes the global scores among the $M(N)(\leq 3^N)$ candidate BN structures; there are at most 3^N DAGs in the case of N variables; for example, if $N = 3$, the eleven quantities are computed and a structure with the largest is chosen.

Note that the second stage does not care about whether each variable is continuous or not. In this paper, we mainly discuss about the performance of the first stage. The number of local scores to be computed can be saved, although it is generally exponential with N. We consider the problem in Section 3.3.

On the other hand, Zhang, Peters, Janzing and Scholkopf [5] proposed a BN structure learning method using conditional independence (CI) tests based on kernel statistics. However, for the CI test that is close to the Hilbert–Schmidt information criterion (HSIC), it is very hard to simulate the null distribution. They only proposed to approximate it by a Gamma distribution, but no consistency, is obtained because the threshold of the statistical test is not correct in practice. Furthermore, for the independence test approach, it often results in conflicting assertions of independence for finite samples. In particular, for small samples, the obtained DAG sometimes contain a directed loop. The Bayesian approach we consider in this paper does not suffer from the inconvenience, because we seek a structure that maximizes the global score [6].

Another contribution of this paper is identifying the border between consistency and non-consistency in learning Bayesian networks. For discrete $\mathcal{X}$, maximizing $Q_X^n(x^n)$ is equivalent to minimizing the description length [1]:

$$-\log Q_X^n(x^n) \approx H^n(x^n) + \frac{\alpha - 1}{2} \log n \,, \tag{3}$$

where $H^n(x^n)$ is the empirical entropy of $x^n \in \mathcal{X}^n$ (we write $A \approx B$ when $|A - B|$ is bounded by a constant) and α is the cardinality of set $\mathcal{X}$. The problem at hand is whether the $\log n$ term is the minimum function of n for ensuring strong consistency. If $\log n$ is replaced by two (AIC), we cannot obtain consistency. We prove that $2(1 + \epsilon) \log \log n$ with $\epsilon > 0$ is the minimum for strong consistency based on the law of iterated logarithms. The same property is known as the Hannan–Quinn principle [7], and similar results have been obtained for autoregression, linear regression [8] and classification [9], among others. The derivation in this paper does not depend on these previous results. The Hannan–Quinn principle will also be applied to continuous variables.

This paper is organized as follows. Section 2.1 introduces the general concept of learning Bayesian network structures based on maximizing the posterior probability, and Section 2.2 discusses the concept of density functions developed by Boris Ryabko [10] and extended by Suzuki [11]. Section 3 presents our contributions: Section 3.1 proves the Hannan–Quinn property in the current problem, and

Section 3.2 proves consistency when continuous variables are present. Section 4 concludes the paper by summarizing the results and states the paper's significance in the field of model selection.

2. Preliminaries

2.1. Learning the Bayesian Structure for Discrete Variables and Its Consistency

We choose w_X, such that $\int w_X(\theta)d\theta = 1$ and $0 \leq \theta(x) \leq 1$ by $w_X(\theta) \propto \prod_{x\in\mathcal{X}} \theta(x)^{-1/2}$, where $\mathcal{X}$ is the set from which X takes its values. Let $\alpha = |\mathcal{X}|$, and let $c_i(x)$ be the frequency of $x \in \mathcal{X}$ in $x^i = (x_1, \cdots, x_i) \in \mathcal{X}^i$, $i = 1, \cdots, n$. It is known that the following quantities satisfies (3) [12]:

$$Q_X^n(x^n) := \prod_{i=1}^{n} \frac{c_{i-1}(x_i) + 1/2}{i - 1 + |\mathcal{X}|/2} = \frac{\Gamma(\alpha/2)\prod_{x\in\mathcal{X}} \Gamma(c_n(x) + 1/2)}{\Gamma(1/2)^\alpha \Gamma(n + \alpha/2)} ,$$

where Γ is the Gamma function, and Stirling's formula $\Gamma(z) = \sqrt{2\pi z}(\frac{z}{e})^z\{1 + O(z^{-1/3})\}$ has been applied. Thus, for $x \in \mathcal{X}$, from the law of large numbers, $c_n(x)/n$ converges to $P(X = x)$ with probability one as $n \to \infty$, such that:

$$-\frac{1}{n}\log Q^n(x^n) \to H(X) := \sum_{x\in\mathcal{X}} -P(X = x)\log P(X = x)$$

with probability one as $n \to \infty$.

Moreover, from the law of large numbers, with probability one as $n \to \infty$,

$$-\frac{1}{n}\log P(X^n = x^n) = \frac{1}{n}\sum_{i=1}^{n}\{-\log P(X = x_i)\} \to E[-\log P(X)] = H(X)$$

(Shannon–McMillan–Breiman [13]). This proves that there exists a Q_X^n (universal measure), such that for any probability P over the finite set $\mathcal{X}$,

$$\frac{1}{n}\log\frac{P^n(x^n)}{Q^n(x^n)} \to 0 \tag{4}$$

with probability one as $n \to \infty$, where we write $P^n(x^n) := P(X^n = x^n)$. The same property holds for:

$$-\log Q_Y^n(y^n) \approx H^n(y^n) + \frac{\beta - 1}{2}\log n , \tag{5}$$

and:

$$-\log Q_{XY}^n(x^n, y^n) \approx H^n(x^n, y^n) + \frac{\alpha\beta - 1}{2}\log n , \tag{6}$$

where $\beta = |\mathcal{Y}|$, $H^n(y^n) = \sum_{y\in\mathcal{Y}} -c_n(y)\log\frac{c_n(y)}{n}$ and $H^n(x^n, y^n) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} -c_n(x, y)\log\frac{c_n(x,y)}{n}$ are the empirical entropies of $y^n \in \mathcal{Y}^n$ and $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, and $c_n(y)$ and $c_n(x, y)$ are the numbers of occurrences of $y \in \mathcal{Y}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in $y^n = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ and $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, respectively.

Thus, we have:

$$J^n(x^n, y^n) := \frac{1}{n}\log\frac{Q_{XY}(x^n, y^x)}{Q_X(x^n)Q_Y(y^n)} \to I(X, Y) := E\{\frac{P(X, Y)}{P(X)P(Y)}\} .$$

with probability one as $n \to \infty$. However, $X \perp\!\!\!\perp Y$ if and only if $I(X,Y) = 0$. Hence, if $X \not\perp\!\!\!\perp Y$, the value of $J^n(x^n, y^n)$ is positive with probability one as $n \to \infty$. However, how can we detect $X \perp\!\!\!\perp Y$ when $X \perp\!\!\!\perp Y$? $J^n(x^n, y^n)$ cannot be exactly zero with probability one as $n \to \infty$.

However, when X and Y are discrete, the estimation based on $J^n(x^n, y^n)$ is consistent: if $X \perp\!\!\!\perp Y$, the value of $J^n(x^n, y^n)$ is not greater than zero with probability one as $n \to \infty$. For example, the decision based on (1) is strongly consistent because the values of $\frac{1}{n}\log p$ and $\frac{1}{n}\log(1-p)$ are negligible for large n, and asymptotically, (1) is equivalent to $J^n(x^n, y^n, z^n) \leq 0$.

In Section 3.1, we provide a stronger result of consistency and a more intuitive and elegant proof.

In general, if N variables exist ($N \geq 2$), we must consider two cases: $D(P^*||P) > 0$ and $D(P^*||P) = 0$, where P^* and P are the probabilities based on the correct and estimated factorizations and $D(P^*||P)$ denotes the Kullback–Leibler divergence between P^* and P. If $N = 2$, then:

$$D(P^*||P) := \sum_x \sum_y P^*(x,y) \log \frac{P^*(x,y)}{P(x,y)} > 0$$

if and only if $X \not\perp\!\!\!\perp Y$ in P^* and $X \perp\!\!\!\perp Y$ in P.

The same property holds for three variables X, Y, Z ($N = 3$):

$$J^n(x^n, y^n, z^n) := \frac{1}{n} \log \frac{Q_{XYZ}(x^n, y^n, z^n) Q_Z^n(z^n)}{Q_{XZ}^n(x^n, y^n) Q_{YZ}^n(y^n, z^n)} \to I(X,Y,Z) := E\{\frac{P(XYZ)P(Z)}{P(XZ)P(YZ)}\}$$

with probability one as $n \to \infty$, and $X \perp\!\!\!\perp Y|Z$ if and only if $I(X,Y,Z) = 0$. Then, we can show $J^n(x^n, y^n, z^n) \leq 0$ if and only if $I(X,Y,Z) = 0$, with probability one as $n \to \infty$ (see Section 3.1). For example, between the seventh and eleventh factorizations, if $J^n(x^n, y^n, z^n) \leq 0$ and $J^n(x^n, y^n, z^n) > 0$, then we choose the seventh and eleventh, respectively. In fact,

$$p_7 \frac{Q_{XZ}^n(x^n, z^n) Q_{YZ}^n(y^n, z^n)}{Q_Z^n(z^n)} \geq p_{11} Q_{XYZ}^n(x^n, y^n, z^n) \Longleftrightarrow J^n(x^n, y^n, z^n) \leq 0$$

for large n, because $\frac{1}{n}\log\frac{p_7}{p_{11}}$ diminishes.

Then, the decision is correct with probability one as $n \to \infty$. Similarly, we calculate:

$$-\log Q_Z(z^n) \approx H^n(z^n) + \frac{\gamma - 1}{2}\log n\,,$$

$$-\log Q_{YZ}(y^n, z^n) \approx H^n(y^n, z^n) + \frac{\beta\gamma - 1}{2}\log n\,,$$

$$-\log Q_{ZX}(z^n, x^n) \approx H^n(z^n, x^n) + \frac{\gamma\alpha - 1}{2}\log n\,,$$

and:

$$-\log Q_{XYZ}(x^n, y^n, z^n) \approx H^n(x^n, y^n, z^n) + \frac{\alpha\beta\gamma - 1}{2}\log n\,,$$

where $\gamma = |\mathcal{Z}|$. In general, for N variables, given P and P^*, we have all of the CI statements for each of them, and $D(P^*||P) = 0$ if and only if the CI statements in P imply those in P^*; in other words, P induces an I-map, which is not necessarily minimal.

Note that for any subsets a, b, c of $\{1, \cdots, N\}$, we can construct the estimation $J^n(x^n, y^n, z^n)$, with $X = \{X^{(i)}\}_{i\in a}, Y = \{Y^{(j)}\}_{j\in b}, Z = \{X^{(k)}\}_{k\in c}$, and obtain consistency, *i.e.*, we will have the correct CI statements, where c may be empty.

Table 1 depicts whether $D(P^*||P) > 0$ or $D(P^*||P) = 0$ for each P^* and P. For example, if the factorizations of P^* and P are the fourth and sixth, then $D(P^*||P) = 0$ from the table. In general, $D(P^*||P) = 0$ if and only if P^* is realized using the factorization and an appropriate parameter set for P.

Table 1. Three-variable case: $D(P^*||P) > 0$ or $D(P^*||P) = 0$: "+" and "0" denote $D(P^*||P) > 0$ and $D(P^*||P) = 0$, respectively.

		Estimated P										
		1	2	3	4	5	6	7	8	9	10	11
True P^*	1	*	0	0	0	0	0	0	0	0	0	0
	2	+	*	+	+	+	0	0	+	+	+	0
	3	+	+	*	+	0	+	0	+	+	+	0
	4	+	+	+	*	0	0	+	+	+	+	0
	5	+	+	+	+	*	+	+	+	+	+	0
	6	+	+	+	+	+	*	+	+	+	+	0
	7	+	+	+	+	+	+	*	+	+	+	0
	8	+	+	+	+	+	+	+	*	+	+	0
	9	+	+	+	+	+	+	+	+	*	+	0
	10	+	+	+	+	+	+	+	+	+	*	0
	11	+	+	+	+	+	+	+	+	+	+	*

2.2. Universal Measures for Continuous Variables

In this section, we primarily address continuous variables.

Let $\{A_j\}$ be such that $A_0 = \{\mathcal{X}\}$, and let A_{j+1} be a refinement of A_j. For example, suppose that the random variable X takes values in $\mathcal{X} = [0, 1]$, and we generate a sequence as follows:

$$
\begin{aligned}
A_1 &= \{[0, \tfrac{1}{2}), [\tfrac{1}{2}, 1)\} \\
A_2 &= \{[0, \tfrac{1}{4}), [\tfrac{1}{4}, \tfrac{1}{2}), [\tfrac{1}{2}, \tfrac{3}{4}), [\tfrac{3}{4}, 1)\} \\
&\vdots \\
A_j &= \{[0, 2^{-(j-1)}), [2^{-(j-1)}, 2 \cdot 2^{-(j-1)}), \cdots, [(2^{j-1} - 1)2^{-(j-1)}, 1)\} . \\
&\vdots
\end{aligned}
$$

For each j, we quantize each $x \in [0, 1]$ into the $a \in A_j$, such that $x \in a$. For example, for $j = 2$, $x = 0.4$ is quantized into $a = [\frac{1}{4}, \frac{1}{2}) \in A_2$. Let λ be the Lebesgue measure (width of the interval). For example, $\lambda([\frac{1}{4}, \frac{1}{2})) = \frac{1}{4}$ and $\lambda(\{\frac{1}{2}\}) = 0$.

Note that each A_j is a finite set. Therefore, we can construct a universal measure Q_j^n w.r.t. a finite set A_j for each j. Given $x^n = (x_1, \cdots, x_n) \in [0, 1]^n$, we obtain a quantized sequence $(a_1^{(j)}, \cdots, a_n^{(j)}) \in A_j^n$ for each j and use it to compute the quantity:

$$g_j^n(x^n) := \frac{Q_j^n(a_1^{(j)}, \cdots, a_n^{(j)})}{\lambda(a_1^{(j)}) \cdots \lambda(a_n^{(j)})}$$

for each j. If we prepare a sequence of positive reals $w_1, w_2, \cdots$, such that $\sum_j w_j = 1$ and $w_j > 0$, we can compute the quantity:

$$g_X^n(x^n) := \sum_{j=1}^{\infty} w_j g_j^n(x^n) \, .$$

Moreover, let f_X be the true density function and $f_j(x) := P(X \in a)/\lambda(a)$ for $a \in A_j$ and $j = 1, 2, \cdots$ if $x \in a$. We may consider f_j to be an approximated density function assuming the quantization sequence $\{A_j\}$ (Figure 2). For the given x^n, we define $f_X^n(x^n) = f_X(x_1) \cdots f_X(x_n)$ and $f_j^n(x^n) := f_j(x_1) \cdots f_j(x_n)$.

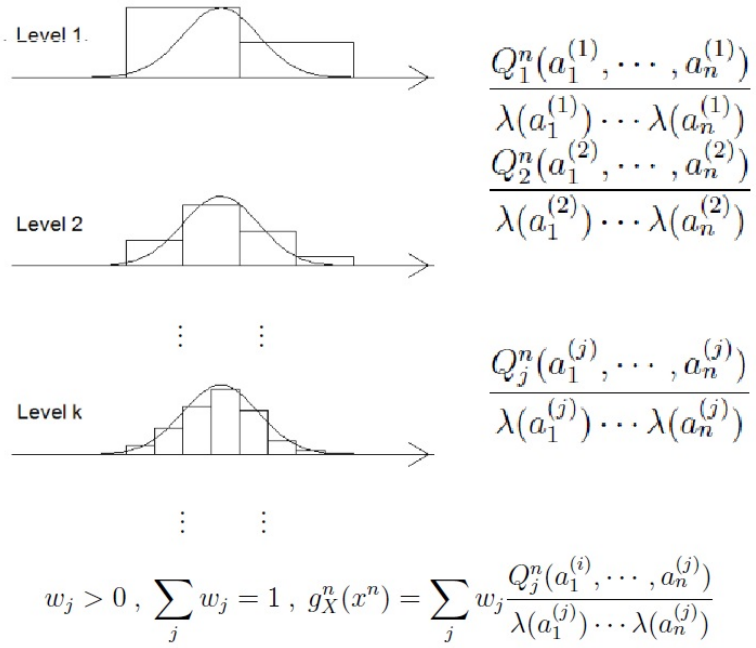

Figure 2. Quantization at level k: $x^n = (x_1, \cdots, x_n) \mapsto (a_1^{(j)}, \cdots, a_n^{(j)})$

Thus, we have the following proposition, which is a continuous version of the universality (4) that was proven in Section 2.1.

Proposition 1 ([10]). *For any density function f, such that $D(f_X||f_j) \to 0$ as $j \to \infty$,*

$$\frac{1}{n} \log \frac{f_X^n(x^n)}{g_X^n(x^n)} \to 0$$

as $n \to \infty$ with probability one, where $D(f_X||f_j)$ is the Kullback–Leibler divergence between f_X and f_j.

The same concept is applied to the case where no density function exists [11] in the usual sense (w.r.t. the Lebesgue measure λ). For example, suppose that we wish to estimate a distribution over the positive integers $\mathbb{N}$. Apparently, $\mathbb{N}$ is not a finite set and has no density function. We consider the quantization sequence $\{B_k\}$: $B_0 = \{\mathbb{N}\}$, $B_1 := \{\{1\}, \{2, 3, \cdots\}\}$, $B_2 := \{\{1\}, \{2\}, \{3, 4, \cdots\}\}$, ..., $B_k := \{\{1\}, \{2\}, \cdots, \{k\}, \{k+1, k+2, \cdots\}\}$, ...

For each k, we quantize each $y \in \mathbb{N}$ into a $b \in B_k$, such that $y \in b$. For example, for $k = 2$, $y = 4$ is quantized into $b = \{3, 4, \cdots\} \in B_2$. Let η be a measure, such that:

$$\eta(\{k\}) = \frac{1}{k} - \frac{1}{k+1}\ ,\ k \in \mathbb{N}\ .$$

The measure $\eta(a)$ for closed interval a gives:

$$\eta(a) = \sum_{k \in a} \eta(\{k\}) = \sum_{k \in a} (\frac{1}{k} - \frac{1}{k+1}) = \frac{1}{k_{min}} - \frac{1}{k_{max}}$$

if k_{min} and k_{max} are the minimum and maximum integers in a, and evaluates each bin width in a nonstandard way. For example, $\eta(\{2\}) = \frac{1}{6}$ and $\eta(\{3, 4\}) = \frac{2}{15}$. For multiple variables, we compute the measure by:

$$\eta(\{j\}, \{k\}) = (\frac{1}{j} - \frac{1}{j+1})(\frac{1}{k} - \frac{1}{k+1})\ .$$

Note that each B_k is a finite set, and we construct a universal measure Q_k^n w.r.t. a finite set B_k for each k. Given $y^n = (y_1, \cdots, y_n) \in \mathbb{N}^n$, we obtain a quantized sequence $(b_1^{(k)}, \cdots, b_n^{(k)}) \in B_k^n$ for each k, such that we can compute the quantity:

$$g_k^n(y^n) := \frac{Q_k^n(b_1^{(k)}, \cdots, b_n^{(k)})}{\eta(b_1^{(k)}) \cdots \eta(b_n^{(k)})}$$

for each k. If we prepare a sequence of positive reals $w_1, w_2, \cdots$, such that $\sum_k w_k = 1$ and $w_k > 0$, we can compute the quantity $g_Y^n(y^n) := \sum_{k=1}^{\infty} w_k g_k^n(y^n)$. In this case, $f_Y(y) = \frac{P(Y = y)}{\eta(\{y\})}$ for $y \in \mathbb{N}$ ($f(y)$ with $y \notin \mathbb{N}$ may take any arbitrary value) is considered to be a generalized density function (w.r.t. the measure η).

In general, if $\eta(D) = 0$ implies $P(Y \in D) = 0$ for the Borel sets (the Borel sets w.r.t. $\mathbb{R}$ being the set consisting of the sets generated via a countable number of union, intersection and set difference from the closed intervals of $\mathbb{R}$ [2]), we state that P is absolutely continuous w.r.t. η and that there exists a density function w.r.t. η (Radon–Nikodym [2]).

The following proposition addresses generalized densities and eliminates the condition $D(f_Y||f_j) \to 0$ as $j \to \infty$ in Proposition 1.

Proposition 2 ([11]). *For any generalized density function f_Y,*

$$\frac{1}{n} \log \frac{f_Y^n(y^n)}{g_Y^n(y^n)} \to 0$$

as $n \to \infty$ with probability one.

Proposition 1 assumes a specific quantization sequence, such as $\{A_n\}$. The universality holds for the densities that satisfy $D(f_X||f_k) \to \infty$ as $k \to \infty$ [10]. However, in the proof of Proposition 2, a universal quantization, such that $D(f_X||f_k) \to 0$ as $k \to \infty$ for any density f_X, was constructed [11].

3. Contributions

3.1. The Hannan and Quinn Principle

We know that $H^n(x^n) + H^n(y^n) - H^n(x^n, y^n)$ is at most $\frac{(\alpha-1)(\beta-1)}{2} \log n$ with probability one as $n \to \infty$ when $X \perp\!\!\!\perp Y$ because the decision based on (1) is strongly consistent.

In this section, we prove a stronger result: let:

$$I^n(x^n, y^n, z^n) := H^n(x^n, z^n) + H^n(y^n, z^n) - H^n(x^n, y^n, z^n) - H^n(z^n) \, .$$

We show that the quantity $I^n(x^n, y^n, z^n)$ is at most $(\alpha - 1)(\beta - 1)\gamma \log\log n$ rather than $\frac{1}{2}(\alpha - 1)(\beta - 1)\gamma \log n$, when $X \perp\!\!\!\perp Y|Z$:

Theorem 1. *If $X \perp\!\!\!\perp Y|Z$:*

$$I^n(x^n, y^n, z^n) \leq (1 + \epsilon)(\alpha - 1)(\beta - 1)\gamma \log\log n \tag{7}$$

with probability one as $n \to \infty$ for any $\epsilon > 0$.

In order to show the claim, we approximate $I^n(x^n, y^n, z^n)$ by $\sum_{z \in \mathcal{Z}} I(z)$ with $I(z) = \frac{1}{2} \sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} r_{i,j}^2$, where $r_{i,j}$, $i = 1, \cdots, \alpha - 1, j = 1, \cdots, \beta - 1$, are mutually independent random variables with mean zero and variance $\sigma_{i,j}^2$, such that:

$$\sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} \sigma_{i,j}^2 = (\alpha - 1)(\beta - 1) \, .$$

Then, from the law of iterated logarithms below (Lemma 1) [2], it will be proven that $r_{i,j}^2$ is almost surely upper-bounded by $2(1 + \epsilon)\sigma_{i,j}^2 \log\log n$ for any $\epsilon > 0$ and each $z \in \mathcal{Z}$, which implies Theorem 1 because:

$$\begin{aligned} I^n(x^n, y^n, z^n) &\approx \sum_z I(z) = \gamma \cdot \frac{1}{2} \sum_i \sum_j r_{i,j}^2 \\ &\leq \gamma \cdot \frac{1}{2} \sum_i \sum_j 2(1 + \epsilon)\sigma_{i,j}^2 \log\log n \\ &= (1 + \epsilon)(\alpha - 1)(\beta - 1)\gamma \log\log n \end{aligned}$$

(see the Appendix for the details of the derivation).

Lemma 1 ([2])**.** *Let $\{U_k\}_{k=1}^n$ be random variables that obey an identical distribution with zero mean and unit variance, and $S_n := \sum_{k=1}^n U_k$. Then, with probability one,*

$$\limsup_{n \to \infty} \frac{S_n}{\sqrt{2n \log n \log n}} = 1 \, .$$

Theorem 1 implies the strong consistency of the decision based on (1). However, a stronger statement can be obtained:

Theorem 2. *We define* $R^n_Z(z^n)$, $R^n_{XZ}(x^n, z^n)$, $R^n_{YZ}(y^n, z^n)$ *and* $R^n_{XYZ}(x^n, y^n, z^n)$ *by:*

$$-\log R^n_Z(z^n) = H^n(z^n) + (1+\epsilon)(\gamma - 1)\log\log n \ ,$$

$$-\log R^n_{XZ}(x^n, z^n) = H^n(x^n, z^n) + (1+\epsilon)(\beta\gamma - 1)\log\log n \ ,$$

$$-\log R^n_{YZ}(y^n, z^n) = H^n(y^n, z^n) + (1+\epsilon)(\beta\gamma - 1)\log\log n \ ,$$

and:

$$-\log R^n_{XYZ}(x^n, y^n, z^n) = H^n(x^n, y^n, z^n) + (1+\epsilon)(\alpha\beta\gamma - 1)\log\log n \ .$$

Then, the decision based on:

$$R^n_{XZ}(x^n, z^n)R^n_{YZ}(y^n, z^n) \geq R^n_{XYZ}(x^n, y^n, z^n)R^n_Z(z^n) \Longleftrightarrow X \perp\!\!\!\perp Y|Z$$

is strongly consistent.

Proof. We note two properties:

1. $R^n_{XZ}(x^n, z^n)R^n_{YZ}(y^n, z^n) \geq R^n_{XYZ}(x^n, y^n, z^n)R^n_Z(z^n)$ is equivalent to (7); and
2. $\lim_{n\to\infty} \frac{1}{n} \log \frac{R^n_{XYZ}(x^n, y^n, z^n)R^n_Z(z^n)}{R^n_{XZ}(x^n, z^n)R^n_{YZ}(x^n, z^n)} = \lim_{n\to\infty} \frac{1}{n} \log \frac{Q^n_{XYZ}(x^n, y^n, z^n)Q^n_Z(z^n)}{Q^n_{XZ}(x^n, z^n)Q^n_{YZ}(x^n, z^n)} \to I(X, Y, Z)$

If $X \perp\!\!\!\perp Y|Z$, then from Theorem 1 and the first property, we have $R^n_{XZ}(x^n, z^n)R^n_{YZ}(y^n, z^n) \geq R^n_{XYZ}(x^n, y^n, z^n)R^n_Z(z^n)$ almost surely. If $R^n_{XZ}(x^n, z^n)R^n_{YZ}(y^n, z^n) \geq R^n_{XYZ}(x^n, y^n, z^n)R^n_Z(z^n)$ almost surely holds, then the value in the second property should be no greater than zero, which means that $X \perp\!\!\!\perp Y|Z$. This completes the proof. □

Theorem 2 is related to the Hannan and Quinn theorem [7] for model selection. To obtain strong consistency, they proved that $\log\log n$ rather than $\frac{1}{2}\log n$ is sufficient for the penalty terms of autoregressive model selection. Recently, several authors have proven this in other settings, such as classification [9] and linear regression [8].

3.2. Consistency for Continuous Variables

Suppose that we wish to estimate the distribution over $[0, 1] \times \mathbb{N}$ in Section 2.2. The set $[0, 1] \times \mathbb{N}$ is not a finite set and has no density function.

Because $A_j \times B_k$ is a finite set, we can construct a universal measure $Q^n_{j,k}$ for $A_j \times B_k$:

$$g^n_{jk}(x^n, y^n) := \frac{Q^n_{j,k}(a^{(j)}_1, \cdots, a^{(j)}_n, b^{(k)}_1, \cdots, b^{(k)}_n)}{\lambda(a^{(j)}_1)\cdots\lambda(a^{(j)}_n)\eta(b^{(k)}_1)\cdots\eta(b^{(k)}_n)}.$$

If we prepare the sequence such that $\sum_{jk}\omega_{jk} = 1$, $\omega_{jk} > 0$, we obtain the quantity:

$$g^n_{XY}(x^n, y^n) := \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} w_{j,k}g^n_{jk}(x^n, y^n) \ .$$

In this case, the (generalized) density function is obtained via:

$$f_{XY}(x, y) = \frac{F_X(x|y)}{dx} \cdot \frac{P(Y = y)}{\eta(\{y\})}$$

where $y \in \mathbb{N}$ (f_{XY} takes arbitrary values for $x \notin [0,1]$ and $y \notin \mathbb{N}$), where $F_X(\cdot|y)$ is the conditional distribution function of X given $Y = y$.

In general, we have the following result:

Proposition 3. *For any generalized density function f:*

$$\frac{1}{n} \log \frac{f_{XY}^n(x^n, y^n)}{g_{XY}^n(x^n, y^n)} \to 0$$

as $n \to \infty$ with probability one.

The measures $g_X^n(x^n)$ and $g_{XY}^n(x^n, y^n)$ are computed using (A) and (B) of Algorithm 1, where the value of K is the number of quantizations, and $\hat{g}_X^n(x^n)$ and $\hat{g}_{XY}^n(x^n, y^n)$ denote the approximated scores using finite quantization of level K.

Algorithm 1 Calculating g^n.

(**A**) Input $x^n \in A^n$, Output $\hat{g}_X^n(x^n)$

1. For each $k = 1, \cdots, K$, $g_k^n(x^n) := 0$
2. For each $k = 1, \cdots, K$ and each $a \in A_k$, $c_k(a) := 0$
3. For each $i = 1, \cdots, n$,
   (a) $A_0 = \mathcal{X}$, $a_i^{(0)} = x_i$
   (b) for each $k = 1, \cdots, K$
      i. Find $a_i^{(k)} \in A_k$ from $a_i^{(k-1)} \in A_{k-1}$
      ii. $\log g_k^n(x^n) := \log g_k^n(x^n) + \log \frac{c_{i,k}(a_i^{(k)}) + 1/2}{i - 1 + |A_k|/2} - \log(\eta_X(a_i^{(k)}))$
      iii. $c_{i,k}(a_i^{(k)}) := c_{i,k}(a_i^{(k)}) + 1$
4. $\hat{g}_X^n(x^n) := \sum_{k=1}^K \frac{1}{K} g_k^n(x^n)$

(**B**) Input $x^n \in A^n$ and $y^n \in B^n$, Output $\hat{g}_{XY}^n(x^n, y^n)$

1. For each $j, k = 1, \cdots, K$, $g_{j,k}^n(x^n, y^n) := 0$
2. For each $j, k = 1, \cdots, K$ and each $a \in A_j$ and $b \in B_k$, $c_{j,k}(a, b) := 0$
3. For each $i = 1, \cdots, n$
   (a) $A_0 = \mathcal{X}$, $B_0 = \mathcal{Y}$, $a_i^{(0)} = x_i$, $b_i^{(0)} = y_i$
   (b) for each $j, k = 1, \cdots, K$
      i. Find $a_i^{(j)} \in A_j$ and $b_i^{(k)} \in B_k$ from $a_i^{(j-1)} \in A_{j-1}$ and $b_i^{(k-1)} \in B_{k-1}$
      ii. $\log g_{j,k}^n(x^n, y^n) := \log g_{j,k}^n(x^n, y^n) + \log \frac{c_{i,j,k}(a_i^{(j)}, b_i^{(k)}) + 1/2}{i - 1 + |A_j||B_k|/2} - \log(\eta_X(a_i^{(j)}) \eta_Y(b_i^{(k)}))$
      iii. $c_{i,j,k}(a_i^{(j)}, b_i^{(k)}) := c_{i,j,k}(a_i^{(j)}, b_i^{(k)}) + 1$
4. $\hat{g}_{XY}^n(x^n, y^n) := \sum_{j=1}^K \sum_{k=1}^K \frac{1}{K^2} g_{j,k}^n(x^n, y^n)$

Propositions 1–3 are obtained for large K. However, we can prepare only a finite number of quantizations. Furthermore, if n is small, then the number of examples that each bin contains is small, and we cannot estimate the histogram well. Therefore, given n, K must be moderately sized, and we recommend to set $K = \frac{1}{m} \log n$ because the number of examples contained in a bin decreases

exponentially with increasing depth, where m is the number of variables in the local score. For example, $m = 1$ and $m = 2$ for (A) and (B), respectively. Algorithm 1 (A)(B) of do not guarantee anything for the theoretical property assured in Proposition 3 and Theorems 3–5 for finite K, however, as K grows, consistency holds.

In Step 3(a) of Algorithm 1(A)(B), we calculate $a_i^{(k)}$ from $a_i^{(k-1)}$ and not from x_i, which means that the computational time required to obtain $(a_i^{(1)}, \cdots, a_i^{(K)})$ from x_i is $O(K)$. Thus, the total computation times of Algorithm 1 (A)(B) are at most $O(nK)$.

In Step 3(b) of Algorithm 1(A), we compute for $i = 1, \cdots, n$ and $k = 1, \cdots, K$:

$$\log \frac{g_k^i(x^i)}{g_k^{i-1}(x^{i-1})} = \log \frac{Q_k^i(a_1^{(k)}, \cdots, a_i^{(k)})}{Q_k^{i-1}(a_1^{(k)}, \cdots, a_{i-1}^{(k)})} - \log \eta_X(a_i^{(k)})$$

if x_i is quantized into $a_i^{(k)} \in A_k$, $i = 1, \cdots, n$.

For the memory requirements, we require exponential orders of K. However, because we set $K = \frac{1}{m} \log n$, the computational time and memory requirements are at most $O(n \log n)$ and $O(n)$ for Algorithm 1(A)(B).

Based on the same notion, we can construct $g_Z^n(z^n)$, $g_{XZ}^n(x^n, z^n)$, $g_{YZ}(y^n.z^n)$, $g_{XYZ}^n(x^n, y^n, z^n)$ from examples $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$ and $z^n \in \mathcal{Z}^n$, and Propositions 2 and 3 hold for three variables.

Theorem 3. *With probability one as $n \to \infty$:*

$$\frac{1}{n} \log \frac{g_{XYZ}^n(x^n, y^n, z^n) g_Z(z^n)}{g_{XY}^n(x^n, z^n) g_{YZ}^n(y^n, z^n)} \to I(X, Y, Z) \tag{8}$$

Proof. From Propositions 2 and 3 for two and three variables and the law of large numbers, we have:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{g_{XYZ}^n(x^n, y^n, z^n) g_Z^n(z^n)}{g_{XZ}^n(x^n, z^n) g_{YZ}^n(y^n, z^n)} = \lim_{n\to\infty} \frac{1}{n} \log \frac{f_{XYZ}^n(x^n, y^n, z^n) f_Z^n(z^n)}{f_{XZ}^n(x^n, z^n) f_{YZ}^n(y^n, z^n)}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \{\log \frac{f_{XYZ}(x_i, y_i, z_i) f_Z(z_i)}{f_{XZ}(x_i, z_i) f_{YZ}(y_i, z_i)}\} = E \log \frac{f_{XYZ}(X, Y, Z)}{f_{XZ}(X, Z) f_{YZ}(Y, Z)} = I(X, Y, Z)$$

with probability one, which completes the proof. □

From the discussion in Section 2.1, even when more than two variables are present, if $D(P^*||P) > 0$, we can choose P^* rather than P with probability one as $n \to \infty$.

Now, we prove that the continuous counterpart of the decision based on (1) is strongly consistent:

Theorem 4. *With probability one as $n \to \infty$:*

$$X \perp\!\!\!\perp Y|Z \Longleftrightarrow p g_{XZ}^n(x^n, z^n) g_{YZ}^n(y^n, z^n) \geq (1-p) g_{XYZ}^n(x^n, y^n, z^n) g_Z^n(z^n) \,, \tag{9}$$

where p is the prior probability of $X \perp\!\!\!\perp Y|Z$.

Proof: Suppose that $X \not\perp\!\!\!\perp Y|Z$. Then, the conditional mutual information between X and Y given Z is positive, and from Theorem 3, the estimator converges to a positive value with probability one as $n \to \infty$; thus, $p g_{XZ}^n(x^n, z^n) g_{YZ}^n(y^n, z^n) \geq (1-p) g_{XYZ}^n(x^n, y^n, z^n) g_Z^n(z^n)$ holds almost surely. Suppose that $X \perp\!\!\!\perp Y|Z$. The discrete variables X and Y are conditionally independent given Z if and only if:

$$c Q_{XZ}^n(x^n, z^n) Q_{YZ}^n(y^n, z^n) \geq (1-c) Q_{XYZ}^n(x^n, y^n, z^n) Q_Z^n(z^n)$$

with probability one as $n \to \infty$ for any constant $0 < c < 1$, even if c does not coincide with the prior probability p. If X, Y and Z are continuous, we quantize x^n, y^n and z^n into $(a_1^{(j)}, \cdots, a_n^{(j)})$, $(b_1^{(k)}, \cdots, b_n^{(k)})$ and $(c_1^{(l)}, \cdots, c_n^{(l)})$. Thus, for each j, k and l, we have:

$$\begin{aligned} & pw_{jl}w_{kl}Q_{jl}^n(a_1^{(j)}, \cdots, a_n^{(j)}, c_1^{(l)}, \cdots, c_n^{(l)})Q_{kl}^n(b_1^{(k)}, \cdots, b_n^{(k)}, c_1^{(l)}, \cdots, c_n^{(l)}) \\ \geq & (1-p)w_{jkl}w_l Q_{jkl}^n(a_1^{(j)}, \cdots, a_n^{(j)}, b_1^{(k)}, \cdots, b_n^{(k)}, c_1^{(l)}, \cdots, c_n^{(l)})Q_n^{(l)}(c_1^{(l)}, \cdots, c_n^{(l)}) \end{aligned}$$

with probability one as $n \to \infty$. Thus, if we divide both sides by:

$$\eta_X(a_1^{(j)}) \cdots \eta_X(a_n^{(j)})\eta_Y(b_1^{(k)}) \cdots \eta_Y(b_n^{(k)})\eta_Z(c_1^{(l)}) \cdots \eta_Z(c_n^{(l)})$$

and take summations of both sides over $j, k.l = 1, 2, \cdots$, we have:

$$pg_{XZ}^n(x^n, z^n)g_{YZ}^n(y^n, z^n) \geq (1-p)g_{XYZ}^n(x^n, y^n, z^n)g_Z^n(z^n)$$

with probability one, where we have assumed $w_{j,k,l} > 0 \Longrightarrow w_{jl}, w_{kl} > 0$ because of $K = \frac{1}{m}\log n$, which completes the proof.

Note that even if either X or Y is discrete, the same conclusion will be obtained. The generalized density functions cover the discrete distributions as a special case.

From the discussion in Section 2.1, even when more than two variables are present, if $D(P^*||P) = 0$, we can choose P^* rather than P with probability one as $n \to \infty$.

Let $h_Z^n(z^n)$,$h_{XZ}^n(x^n, z^n)$, $h_{YZ}^n(y^n, z^n)$ and $h_{XYZ}^n(x^n, y^n, z^n)$ take the same values of $g_Z^n(z^n)$, $g_{XZ}^n(x^n, z^n)$, $g_{YZ}^n(y^n, z^n)$ and $g_{XYZ}^n(x^n, y^n, z^n)$, except that the $\log n$ terms in $-\log Q_Z^n(z^n)$, $-\log Q_{XZ}^n(x^n, z^n)$,$-\log Q_{YZ}^n(y^n, z^n)$ and $-\log Q_{XYZ}^n(x^n, y^n, z^n)$ are replaced by $2(1+\epsilon)\log\log n$, respectively, where $\epsilon > 0$ is arbitrary. Then, we obtain the final result:

Theorem 5. *With probability one as $n \to \infty$:*

$$ph_{XZ}^n(x^n, z^n)h_{YZ}^n(y^n, z^n) \geq (1-p)h_{XYZ}^n(x^n, y^n, z^n)h_Z^n(z^n) \Longleftrightarrow X \perp\!\!\!\perp Y|Z. \tag{10}$$

This paper focuses on the theoretical aspects of the BN structure learning, in particular for consistency when continuous variables are present. For the details of the practical matters we deal with in this section, see the conference paper [14].

3.3. The Number of Local Scores to be Computed

We refer the conditional independence (CF) score w.r.t. X and Y given Z to the left of (8). Suppose we follow the fastest Bayesian network structure learning due to [6]: let $Pa(X, V)$ be the optimal parent set of $X \in V$ contained in $V - \{X\}$ for $V \subseteq U := \{1, \cdots, N\}$ and $S(X, V)$ its local score. Then, we can obtain:

$$T(V) := \max_{x \in V}\{S(X, V) + T(V - \{X\})\}$$

For each $V \subseteq U$, the sinks:

$$X_N = \mathrm{argmax}_{X \in U}T(U)\ ,\ X_{N-1} = \mathrm{argmax}_{X \in U - \{X_N\}}T(U - \{X_N\})\ ,\ \cdots\ ,$$

and the parent sets:

$$Pa(X_N, U)\ ,\ P(X_{N-1}, U - \{X_N\})\ ,\ \cdots\ ,\ \{\}\ .$$

For each fixed pair (X, V), maximizing the local score $\frac{1}{n}\log\frac{g_{W+\{X\}}}{g_W}$ and maximizing the CF score $\frac{1}{n}\log\frac{g_{V-\{X\}}g_{W+\{X\}}}{g_V g_W}$ w.r.t. $V - \{X\}$ and $W + \{X\}$, given W are equivalent. In other words,

$$\frac{1}{n}\log\frac{g_{W+\{X\}}}{g_W} \leq \frac{1}{n}\log\frac{g_{W'+\{X\}}}{g_{W'}} \Longleftrightarrow \frac{1}{n}\log\frac{g_{V-\{X\}}g_{W'+\{X\}}}{g_V g_W} \leq \frac{1}{n}\log\frac{g_{V-\{X\}}g_{W'+\{X\}}}{g_V g_{W'}}$$

for $W, W' \subseteq V - \{X\}$.

On the other hand, from [15,16], we know that the relationship between the complexity term and the likelihood term gives tight bounds on the maximum number of parents in the optimal BN for any given dataset. In particular, the number of elements in each parent set $Pa(X, V)$ is at most $O(\log n)$ for $X \in V$ and $V \subseteq U$. Hence, the number for computing the CF scores is much less than exponential with N.

4. Concluding Remarks

In this paper, we considered the problem of learning a Bayesian network structure from examples and provided two contributions.

First, we found that the $\log n$ terms in the penalty terms of the description length can be replaced by $2(1+\epsilon)\log\log n$ to obtain strong consistency, where the derivation is based on the law of iterated logarithms. We claim that the Hannan and Quinn principle [7] is applicable to this problem.

Second, we constructed an extended version of the score function for finding a Bayesian network structure with the maximum posterior probability and proved that the decision is strongly consistent even when continuous variables are present. Thus far, consistency has been obtained only for discrete variables, and many authors have been seeking consistency when continuous variables are present.

Consistency has been proven in many model selection methods that maximize the posterior probability or, equivalently, minimize the description length [1]. However, almost all such methods assume that the variables are either discrete or that the variables obey Gaussian distributions. This paper proposed an extended version of the MDL/Bayesian principle without assuming such constraints and proved its strong consistency in a precise manner, which we believe provides a substantial contribution to the statistics and machine learning communities.

Appendix: Proof of Theorem 1

Hereafter, we write $P(X = x|Z = z)$ and $P(Y = y|Z = z)$ simply as $P(x|z)$ and $P(y|z)$, respectively, for $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. We find that the empirical mutual information:

$$
\begin{aligned}
& I^n(x^n, y^n, z^n) \\
= & \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} c_n(x,y,z)\log\frac{c_n(x,y,z)c_n(z)}{c_n(x,z)c_n(y,z)} \\
= & \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} c_n(z)P(x|z)P(y|z)\cdot\frac{c_n(x,y,z)}{c_n(z)P(x|z)P(y|z)}\log\frac{c_n(x,y,z)}{c_n(z)P(x|z)P(y|z)} \\
& -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} c_n(z)P(x|z)\cdot\frac{c_n(x,y,z)}{c_n(z)P(x|z)}\log\frac{c_n(x,z)}{c_n(z)P(x|z)} \\
& -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} c_n(z)P(y|z)\cdot\frac{c_n(x,y,z)}{c_n(z)P(y|z)}\log\frac{c_n(y,z)}{c_n(z)P(y|z)} \\
= & \sum_{z\in\mathcal{Z}}\{ \\
& \sum_x\sum_y c_n(z)P(x|z)P(y|z)\cdot\frac{c_n(x,y,z)}{c_n(z)P(x|z)P(y|z)}\log\frac{c_n(x,y,z)}{c_n(z)P(x|z)P(y|z)} \qquad (11) \\
& -\sum_x c_n(z)P(x|z)\cdot\frac{c_n(x,z)}{c_n(z)P(x|z)}\log\frac{c_n(x,z)}{c_n(z)P(x|z)} \qquad (12) \\
& -\sum_y c_n(z)P(y|z)\cdot\frac{c_n(y,z)}{c_n(z)P(y|z)}\log\frac{c_n(y,z)}{c_n(z)P(y|z)} \qquad (13) \\
& \}
\end{aligned}
$$

is approximated by $\sum_{z\in\mathcal{Z}} I(z)$ with:

$$
\begin{aligned}
I(z) := & \sum_x\sum_y \frac{\{c_n(x,y,z) - c_n(z)P(x|z)P(y|z)\}^2}{2c_n(z)P(x|z)P(y|z)} \\
& -\sum_x \frac{\{c_n(x,z) - c_n(z)P(x|z)\}^2}{2c_n(z)P(x|z)} - \sum_y \frac{\{c_n(y,z) - c_n(z)P(y|z)\}^2}{2c_n(z)P(y|z)}
\end{aligned}
$$

where the difference between them is zero with probability one as $n \to \infty$, and $(1+t)\log(1+t) = t + t^2/2 - t^3/\{6[1+\delta(t)t]^2\}$ with $0 < \delta(t) < 1$ and:

$$
t = \frac{c_n(x,y,z)}{c_n(z)P(x|z)P(y|z)} - 1, \frac{c_n(x,z)}{c_n(z)P(x|z)} - 1, \frac{c_n(y,z)}{c_n(z)P(y|z)} - 1
$$

has been applied for (11), (12) and (13), respectively. Furthermore, we derive:

$$
I(z) = \frac{1}{2}\mathrm{trace}({}^tVV) - \frac{1}{2}||{}^tuV||^2 - \frac{1}{2}||Vw||^2 \,, \qquad (14)
$$

where $V = (V_{xy})_{x\in\mathcal{X},y\in\mathcal{Y}}$ with $V_{xy} = \dfrac{c_n(x,y,z) - c_n(z)P(x|z)P(y|z)}{\sqrt{2c_n(z)P(x|z)P(y|z)}}$, and u and v are the column vectors $[\sqrt{P(x|z)}]_{x\in\mathcal{X}}$ and $[\sqrt{P(y|z)}]_{y\in\mathcal{Y}}$, respectively. Hereafter, we arbitrarily fix $z \in \mathcal{Z}$. Let $U =$

$(u[0], u[1], \cdots, u[\alpha-1])$ with $u[0] = u$ and $W = (w[0], w[1], \cdots, w[\beta-1])$ with $w[0] = w$ being eigenvectors of $E_\alpha - [\sqrt{P(x|z)P(x'|z)}]_{x,x' \in \mathcal{X}}$ and $E_\beta - [\sqrt{P(y|z)P(y'|z)}]_{y,y' \in \mathcal{Y}}$, where E_m is the identity matrix of dimension m.

Then, ${}^t uVw = 0$, and for $\tilde{U} = (u[1], \cdots, u[\alpha-1])$ and $\tilde{W} = (w[1], \cdots, w[\beta-1])$, we have:

$$ {}^tUVW = \left[\begin{array}{c} {}^t u \\ {}^t \tilde{U} \end{array} \right] V \left[\begin{array}{cc} w & \tilde{W} \end{array} \right] = \left[\begin{array}{cc} 0 & {}^t uV\tilde{W} \\ {}^t \tilde{U}Vw & {}^t \tilde{U}V\tilde{W} \end{array} \right] . $$

and:

$$ \begin{aligned} {}^t({}^tUVW)({}^tUVW) &= \left[\begin{array}{cc} 0 & {}^t\{{}^t\tilde{U}Vw\} \\ {}^t\{{}^t uV\tilde{W}\} & {}^t\{{}^t\tilde{U}V\tilde{W}\} \end{array} \right] \left[\begin{array}{cc} 0 & {}^t uV\tilde{W} \\ {}^t\tilde{U}Vw & {}^t\tilde{U}V\tilde{W} \end{array} \right] \\ &= \left[\begin{array}{cc} {}^t\{{}^t\tilde{U}Vw\} \cdot {}^t\tilde{U}Vw & {}^t\{{}^t\tilde{U}Vw\}{}^t\tilde{U}V\tilde{W} \\ {}^t\{{}^t\tilde{U}V\tilde{W}\}{}^t\tilde{U}Vw & {}^t\{{}^t uV\tilde{W}\}{}^t uV\tilde{W} + {}^t\{{}^t\tilde{U}V\tilde{W}\}{}^t\tilde{U}V\tilde{W} \end{array} \right] . \end{aligned} $$

If we note that $U{}^tU = {}^tUU = E_\alpha$ and $W{}^tW = {}^tWW = E_\beta$, we obtain:

$$ \mathrm{trace}({}^tVV) = \mathrm{trace}({}^t\{{}^tUVW)({}^tUVW)\} = {}^t\{Vw\}Vw + {}^t\{{}^tuV\}{}^tuV + trace{}^t\{{}^t\tilde{U}V\tilde{W})({}^t\tilde{U}V\tilde{W}\} $$

and find that (14) becomes:

$$ I(z) = \frac{1}{2} trace[{}^t\{{}^t\tilde{U}V\tilde{W}\}{}^t\tilde{U}V\tilde{W}] = \frac{1}{2} \sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} r_{ij}^2 $$

with $r_{ij} := {}^t u[i]Vw[j]$. Then, we can see:

$$ E[2I(z)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \{1 - P(x|z)P(y|z)\} - \sum_{x \in \mathcal{X}} \{1 - P(x|z)\} - \sum_{y \in \mathcal{Y}} \{1 - P(y|z)\} = (\alpha-1)(\beta-1) , \quad (15) $$

and that the $(\alpha-1) \times (\beta-1)$ matrix ${}^t\tilde{U}V\tilde{W}$ consists of mutually independent elements r_{ij} with $i = 1, \cdots, \alpha-1$ and $j = 1, \cdots, \beta-1$: $E[r_{ij}] = 0$, and:

$$ E[r_{ij} r_{i'j'}] = \left\{ \begin{array}{ll} \sigma_{ij}^2, & (i,j) = (i',j') \\ 0, & \text{otherwise} \end{array} \right. , $$

where σ_{ij}^2 is the variance of r_{ij} and the expectation of r_{ij}^2, so that (15) implies:

$$ \sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} \sigma_{ij}^2 = (\alpha-1)(\beta-1) . \quad (16) $$

If we define for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and for $i = 1, \cdots, n$:

$$ Z_{i,j,k} := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} u[i,x] w[y,j] \frac{I(X = x_k, Y = y_k, Z = z_k) - P(x|z)P(y|z)}{\sqrt{P(x|z)P(y|z)}\sigma_{ij}} , $$

where $u[i] = (u[i,x])_{x \in \mathcal{X}}$ and $w[j] = (w[y,j])_{y \in \mathcal{Y}}$, then we can check $E[Z_{i,j,k}] = 0$ and $V[Z_{i,j,k}] = 1$, where expectation E and variance V are with respect to the examples $X^n = x^n$ and $Y^n = y^n$, and $I(A)$ takes one if the event A is true and zero otherwise. We can easily check:

$$ \sum_{k=1}^{n} Z_{i,j,k} = \sqrt{c_n(z)} \frac{r_{ij}}{\sigma_{ij}} . \quad (17) $$

We consider applying the obtained derivation to Lemma 1. From (17), we obtain:

$$1 = \limsup_{n\to\infty} \frac{\sqrt{c_n(z)} \cdot r_{ij}}{\sqrt{2c_n(z)\log\log c_n(z)} \cdot \sigma_{ij}} = \limsup_{n\to\infty} \frac{r_{ij}}{\sigma_{ij}\sqrt{2n\log\log n}}$$

which means that (14) is upper bounded by $(1+\epsilon)(\alpha-1)(\beta-1)\log\log n$ with probability one as $n \to \infty$ for any $\epsilon > 0$, from (16). This completes the proof of Theorem 1.

References

1. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471.
2. Billingsley, P. Probability & Measure, 3rd ed.; Wiley: New York, NY, USA, 1995.
3. Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **2000**, *7*, 601–620.
4. Imoto, S.; Kim, S.; Goto, T.; Aburatani, S.; Tashiro, K.; Kuhara, S.; Miyano, S. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.* **2003**, *1*, 231–252.
5. Zhang, K.; Peters, J.; Janzing, D.; Scholkopf, B. Kernel-based Conditional Independence Test and Application in Causal Discovery. In Proceedings of the 2011 Uncertainty in Artificial Intelligence Conference, Barcelona, Spain, 14–17 July 2011; pp. 804–813.
6. Silander, T.; Myllymaki, P. A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, 13–16 July 2006; pp. 445–452.
7. Hannan, E.J.; Quinn, B.G. The Determination of the Order of an Autoregression. *J. R. Stat. Soc. B* **1979**, *41*, 190–195.
8. Suzuki, J. The Hannan–Quinn Proposition for Linear Regression, *Int. J. Stat. Probab.* **2012**, *1*, 2.
9. Suzuki, J. On Strong Consistency of Model Selection in Classification. *IEEE Trans. Inf. Theory* **2006**, *52*, 4767–4774,
10. Ryabko, B. Compression-based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series, *IEEE Trans. Inform. Theory* **2009**, *55*, 4309–4315, .
11. Suzuki, J. Universal Bayesian Measures, In Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 644–648.
12. Krichevsky, R.E.; Trofimov, V.K. The Performance of Universal Encoding. *IEEE Trans. Inf. Theory* **1981**, *27*, 199–207.
13. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 1995.
14. Suzuki, J. Learning Bayesian Network Structures When Discrete and Continuous Variables Are Present. In Proceedings of the 2014 Workshop on Probabilistic Graphical Models, 17-19 September 2014; Springer Lecture Notes on Artificial Intelligence, Volume 8754; pp. 471–486.
15. Suzuki, J. Learning Bayesian belief networks based on the minimum description length principle: An efficient algorithm using the B&B technique. In Proceedings of the 13th International Conference on Machine Learning (ICML'96), Bari, Italy, 3–6 July 1996; pp. 462–470.

16. De Campos, C.P.; Ji, Q. Efficient Structure Learning of Bayesian Networks using Constraints. *JMLR* **2011**, *12*, 663–689.
17. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
18. Judea, P. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference; Morgan-Kaufmann: San Mateo, CA, USA, 1988.