The Bayesian Chow-Liu Algorithm

Joe Suzuki
Osaka University, Japan
suzuki@math.sci.osaka-u.ac.jp

Abstract

Given data, not knowing the distribution, we wish to construct a forest (Markov graph) relative to which the description length is minimized, connecting edges with larger estimated mutual information of each pair of random variables step by step (the Chow-Liu algorithm) to balance simplicity of the forest and fitness of the data, where the random variables are not to be either discrete or continuous. To this end, we construct a Bayesian measure over the data sequences, and propose the Bayesian estimator of mutual information. The Bayesian measure is partially from (Ryabko 2009), but our version can deal with any random variable even if no density function exists. We show that the estimator is consistent, and it is considered to be more robust than the existing approaches because it does not estimate one specific histogram from data. Numerical experiments demonstrate that the proposed method works efficiently enough to deal with practical problems.

1 Introduction

In many applications of statistical machine learning such as data mining and pattern recognition, we often need to capture the dependencies among random variables from data. The obtained relation can be usually expressed by graphical models such as Markov networks and Bayesian networks (Pearl 1988). However, as the number of random variables increases, it is hard to obtain the exact estimation because its computation increases exponentially.

In this paper, we restrict the distribution of random variables expressed by a Markov graph to a limited form of distributions expressed by a tree (we identify the random variables with the vertexes in the graph). If the distribution is known, such an approximation can be executed via the Chow-Liu algorithm (Chow and Liu, 1968) which continues to connect a pair of vertexes with the largest mutual information if the connection does not make any loop (otherwise, the pair will not be considered for an edge in the future) until no candidate exists. Although the search is done in a top down manner, it is guaranteed that the resulting tree expresses a distribution such that the K-L divergence from the true distribution is minimized (Section 2.1).

In our problem, only the data is available while the true distribution is not known. Given n examples consisting of attribute values, a naive way to construct a distribution expressing a tree is to maximum likelihood estimate the values of mutual information based the examples (Section 2.2). On the other hand, (Suzuki, 1993) considered to minimize the description length rather than the K-L divergence by estimating each mutual information in a Bayesian manner: construct measures R^n_X, R^n_Y, R^n_{XY} over $\mathcal{X}^n$, $\mathcal{Y}^n$, $\mathcal{X}^n \times \mathcal{Y}^n$, where $\mathcal{X}$, $\mathcal{Y}$ are the ranges of random variables X, Y. Then, the Bayesian estimator is expressed by $\dfrac{1}{n}\log\dfrac{R^n_{XY}(x^n,y^n)}{R^n_X(x^n)R^n_Y(y^n)}$ for given examples $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, reflecting simplicity of each forest as well as the likelihood of the examples to the tree (Section 2.3).

The main purpose of this paper is to extend the Chow-Liu algorithm so that it can deal with arbitrary random variables: the existing methods deal with only random variables taking values in finite sets. Suppose that all the variables are continuous and a simultaneous density function exists. Then, constructing a kernel function to which the training data fit may show better performance in some cases. However, in reality, in any data base, some attributes are discrete, and others are continuous. So, assuming only continuous variables would be too restrictive for the multivariate case. We do not assume that the random variables are either discrete or continuous. To this end, we need to extend the notion of description length and Bayesian estimators of mutual information. We require such a Bayesian estimator to be (strongly) consistent, i.e., the estimator should converge to the true one as $n \to \infty$ with probability one.

There are many ways to estimate mutual information. Most conventional approaches were to quantize the ranges $\mathcal{X}, \mathcal{Y}$ for estimation and to increase the number of bins in the histogram as the sample size n grows: (Darbellay and Vajda 1999) considered to update the bins adaptively based on the samples obtained thus far (strong consistency was not proved for the method); (Wang et al 2005) applied a similar idea to estimation of Kullback-Leibler divergence; and recently (Silva and Narayanan 2010) obtained a consistent estimator of mutual information using a similar but more general principle.

We construct a measure g^n over $\mathcal{X}^n$ that is universal in the sense $\dfrac{1}{n}\log\dfrac{f_X^n(x^n)}{g_X^n(x^n)}$ diminishes with probability as one $n \to \infty$ for any f_X^n. The idea is to prepare a nested sequence $\{A_k\}$ of histograms and to estimate the density function $f_k^n(x^n)$ by $g_k^n(x^n)$ for each histogram A_k, assuming that the density function f_X exists for the random variable X. Then, we mix them with weights $\{w_k\}$ such that $\sum_{k=1}^{\infty} w_k = 1, w_k > 0$ to obtain the value $g_X^n(x^n) = \sum_{k=1}^{\infty} w_k g_k^n(x^n)$. For the measure, (Ryabko 2009) proved universality for any f_X such that $h(f_k) \to h(f_X)$ as $k \to \infty$ (Section 3.2).

We extend the universal measure g^n so that the random variables can be either discrete or continuous. The basic idea is to replace the Lebesgue measure λ with another supporting measure η if no density function exists for the random variable X. In particular, if $\mathcal{X}$ is finite, the g_X^n reduces to R_X^n by choosing an appropriate η, as shown in Section 3.3.

The proposed extended Bayesian estimator is expressed by $\dfrac{1}{n}\log\dfrac{g_{XY}^n(x^n,y^n)}{g_X^n(x^n)g_Y^n(y^n)}$, where g_X^n, g_Y^n, g_{XY}^n are universal density functions with respect to $\eta \neq \lambda$. We show that the estimator actually converges to the mutual information with probability one as $n \to \infty$ (Section 4.1).

There are many consistent estimators of mutual information. The proposed Bayesian estimator has several merit over them. For example, by maximizing it for each step, the associate description length assuming the graph (V,E) to be a forest will be minimized among E (Section 4.2).

In the last section (Section 5), we list other merits of the Bayesian estimator and state future works.

2 The Chow-Liu Algorithm

2.1 The Original Version

Let V be a finite set, and E a subset of $\mathcal{E} := \{\{i,j\} \subseteq V | i \neq j\}$. In this paper, we say such a pair (V,E) to be a *graph*. The sequences $(i_0,\cdots,i_m),(i_m,\cdots,i_0) \in V^{m+1}$ are said to be a pair of *paths* connecting $\{i_0,i_m\}$ of length m in (V,E) if $i_0,\cdots,i_m$ are different and $\{i_0,i_1\},\cdots,\{i_{m-1},i_m\} \in E$. We say the graph (V,E) is a *forest* if a pair of paths connecting each element in $\mathcal{E}$ is unique (if it exists), and a *tree* if there exists a unique path connecting each element in $\mathcal{E}$.

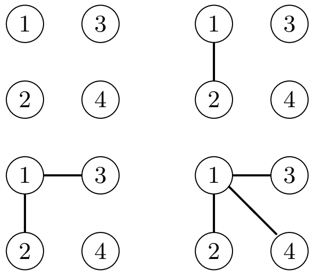
In this paper, we consider the following algorithm for graphs: given $\{w_{i,j}\}_{\{i,j\}\in\mathcal{E}}$ such that $w_{i,j} \geq 0$, $w_{i,j} = w_{j,i}$, it outputs a tree maximizing $\sum_{\{i,j\}\in E} w_{i,j}$ among graphs (V,E) that express trees (Kruskal's algorithm, Aho 1974):

```
1. $\mathcal{E} \leftarrow \{\{i,j\} | i,j \in V, i \neq j\}$
2. $E \leftarrow \{\}$
3. while($\mathcal{E} \neq \phi$) for $\{i,j\} \in \mathcal{E}$ maximizing $w_{i,j}$
    (a) $\mathcal{E} \leftarrow \mathcal{E} \backslash \{\{i,j\}\}$
    (b) $(V, E \cup \{\{i,j\}\})$ is a forest $\Longrightarrow E \leftarrow E \cup \{\{i,j\}\}$
```

If the initial $\mathcal{E}$ is replaced by $\{\{i,j\}|i,j \in V, i \neq j, w_{i,j} > 0\}$, then the resulting graph is a forest rather than a tree (the generalized Kruskal algorithm).

Example 1. Suppose that the values of $\{w_{i,j}\}_{\{i,j\}\in\mathcal{E}}$ are given in the table below. The largest value in the table is twelve for $(i,j) = (1,2)$, so we connect them first. The second largest is ten for $(i,j) = (1,3)$, so we connect them. The third largest is eight for $(i,j) = (2,3)$, but connecting them makes a loop, so we do not connect them. The fourth largest is six for $(i,j) = (1,4)$, so we connect them. But, we cannot connect any further for the rest $(i,j) = (2,4), (3,4)$.

i	1	1	2	1	2	3
j	2	3	3	4	4	4
$w_{i,j}$	12	10	8	6	4	2



Let $X^{(1)}, \cdots, X^{(N)}$ be random variables that take values in finite sets $\mathcal{X}^{(1)}, \cdots, \mathcal{X}^{(N)}$, respectively. Let $P_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$, $P_i(x^{(i)})$, and $P_{i,j}(x^{(i)}, x^{(j)})$ be the probabilities of $(X^{(1)}, \cdots, X^{(N)}) = (x^{(1)}, \cdots, x^{(N)}) \in \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(N)}$, $X^{(i)} = x^{(i)} \in \mathcal{X}^{(i)}$, and $(X^{(i)}, X^{(j)}) = (x^{(i)}, x^{(j)}) \in \mathcal{X}^{(i)} \times \mathcal{X}^{(j)}$, respectively. Also, let $H(i)$, $I(i,j)$, and $H(1, \cdots, N)$ be the entropy of $X^{(i)}$, the mutual information of $\{X^{(i)}, X^{(j)}\}$, and the simultaneous entropy of $\{X^{(1)}, \cdots, X^{(N)}\}$, respectively.

Let $V := \{1, \cdots, N\}$, and $E \subseteq \mathcal{E} = \{\{i,j\} \subseteq V | i \neq j\}$. Assuming (V, E) is a tree, we identify V with $\{X^{(1)}, \cdots, X^{(N)}\}$ to approximate $P_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$by

$$Q_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)}) = \frac{\prod_{\{i,j\}\in E} P_{i,j}(x^{(i)}, x^{(j)})}{\prod_{i\in V} P_i(x^{(i)})^{d_i - 1}}$$

(Dendroid distribution), where $d_i := |\{j \in V | \{i,j\} \in E\}|$[1].

Suppose that we rearrange the indexes $1, \cdots, N$ so that $i \leq j$ if path $(1, \cdots, i, \cdots, j)$ exists. Then, for each $j = 2, \cdots, N$, the i such that $i < j$ and $\{i,j\} \in E$ is unique:

$$\begin{aligned} & Q_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)}) \\ = \; & P_1(x^{(1)}) \prod_{\{i,j\}\in E, i<j} \frac{P_{i,j}(x^{(i)}, x^{(j)})}{P_i(x^{(i)})} . \end{aligned}$$

In 1968, Chow and Liu showed that if we apply Kruskal's algorithm with weights $w_{i,j} := I(i,j)$, then, the resulting tree minimizes the Kullback-Leibler divergence

$$\begin{aligned} & D(P_{1,\cdots,N} || Q_{1,\cdots,N}) \\ = \; & \sum_{i\in V} H(i) - H(1, \cdots, N) - \sum_{\{i,j\}\in E} I(i,j) \end{aligned}$$

depends only on E in the last term.

2.2 The Chow-Liu Algorithm based on ML Estimation

Suppose that $x^n := \{(x_i^{(1)}, \cdots, x_i^{(N)})\}_{i=1}^n \in (\prod_{j=1}^N \mathcal{X}^{(j)})^n$ have been emitted i.i.d. by $P_{1,\cdots,N}$. Let $c_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$, $c_{i,j}(x^{(i)}, x^{(j)})$, and $c_i(x^{(i)})$ be the numbers of occurrences of $(X^{(1)}, \cdots, X^{(N)}) = (x^{(1)}, \cdots, x^{(N)}) \in (\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(N)})$, $(X^{(i)}, X^{(j)}) = (x^{(i)}, x^{(j)}) \in \mathcal{X}^{(i)} \times \mathcal{X}^{(j)}$, and $X^{(i)} = x^{(i)} \in \mathcal{X}^{(i)}$, in x^n, respectively. If we divide them by n, we obtain the values of $\hat{P}_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$, $\hat{P}_{i,j}(x^{(i)}, x^{(j)})$, and $\hat{P}_i(x^{(i)})$, respectively.

If we define $\hat{Q}_{1,\cdots,N}$, $\hat{H}(i)$, $\hat{H}(1, \cdots, N)$, $\hat{I}(i,j)$ in terms of $\hat{P}_{1,\cdots,N}, \hat{P}_{i,j}, \hat{P}_i$ (the maximum likelihood estimators), we obtain a tree minimizing

$$\begin{aligned} & D(\hat{P}_{1,\cdots,N} || \hat{Q}_{1,\cdots,N}) \\ = \; & \sum_{i\in V} \hat{H}(i) - \hat{H}(1, \cdots, n) - \sum_{\{i,j\}\in E} \hat{I}(i,j) \end{aligned}$$

2.3 The Chow-Liu Algorithm based on the MDL

Let $R^n(i)$ and $R^n(i,j)$ be measures over $(\mathcal{X}^{(i)})^n$ such that $\sum R^n(i) \leq 1$ and over $(\mathcal{X}^{(i)} \times \mathcal{X}^{(j)})^n$

[1] $|S|$ denotes the cardinality of set S.

such that $\sum R^n(i,j) \le 1$, respectively. Then, if we define the measure over $(\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(N)})^n$ by

$$R^n(1,\cdots,N|E) := \frac{\prod_{\{i,j\}\in E} R^n(i,j)}{\prod_{i\in V} R^n(i)^{d_i-1}}$$
$$= \prod_{\{i,j\}\in E} \frac{R^n(i,j)}{R^n(i)R^n(j)} \prod_{i\in V} R^n(i),$$

then the Chow-Liu algorithm maximizing the Bayesian estimator

$$J(i,j) \;:=\; \frac{1}{n}\log\frac{R^n(i,j)}{R^n(i)R^n(j)} \qquad (1)$$

in each step minimizes the associated description length (Rissanen 1978)

$$L(x^n|E) := -\log R^n(1,\cdots,N|E)$$
$$= -\sum_{i\in V}\log R^n(i) - \sum_{\{i,j\}\in E}\log\frac{R^n(i,j)}{R^n(i)R^n(j)} \ .$$

However, the E minimizing the description length $L(x^n|E)$ may not converge to the true E as $n \to \infty$. In fact, for example, if the values of $R^n(i)$ and $R^n(i,j)$ are uniform and do not depend on x^n, the quantity $L(x^n|E)$ does not give any information to estimate E, even if we properly specify the prior probability $P(E)$ over the subsets of $\mathcal{E}$.

Hereafter, for simplicity, we assume that the prior probability $P(E)$ is uniform, so that, given x^n, we evaluate (V,E) only by $L(x^n|E)$ rather than by $-\log P(E) + L(x^n|E)$.

On the other hand, if we apply the Krichevsky-Trofimov estimator (Krichevsky and Trofimov, 1981) such as

$$R^n(i) := \frac{\Gamma(n+\alpha^{(i)}a)\Gamma(a)^{\alpha^{(i)}}}{\Gamma(\alpha^{(i)}a)\prod_{x^{(i)}\in\mathcal{X}^{(i)}}\Gamma(c_i[x^{(i)}]+a)} \quad (2)$$

$$R^n(i,j)$$
$$:= \{\Gamma(n+\alpha^{(i)}\alpha^{(j)}a)\Gamma(a)^{\alpha^{(i)}\alpha^{(j)}}\}/\{\Gamma(\alpha^{(i)}\alpha^{(j)}a)$$
$$\prod_{x^{(i)}\in\mathcal{X}^{(i)},x^{(j)}\in\mathcal{X}^{(j)}}\Gamma(c_{i,j}[x^{(i)},x^{(j)}]+a)\} \ ,$$

with parameter $a = 1/2$, we obtain[2]

$$-\log R^n(i) \approx n\hat{H}(i) + \frac{\alpha^{(i)}-1}{2}\log n \ ,$$

$$-\log R^n(i,j) \approx n\hat{H}(i,j) + \frac{\alpha^{(i)}\alpha^{(j)}-1}{2}\log n \ ,$$

$$\log\frac{R^n(i,j)}{R^n(i)R^n(j)} \approx n\hat{I}(i,j) - \frac{(\alpha^{(i)}-1)(\alpha^{(j)}-1)}{2}\log n \ ,$$

and thus

$$L(x^n|E) \approx n\sum_{i\in V}\{\hat{H}(i) + \frac{\alpha^{(i)}-1}{2n}\log n\}$$
$$-n\sum_{\{i,j\}\in E}\{\hat{I}(i,j) - \frac{1}{2n}(\alpha^{(i)}-1)(\alpha^{(j)}-1)\log n\} \ .$$

Since $\frac{1}{n}\log R^n(i) \to H(i)$ and $\frac{1}{n}\log R^n(i,j) \to H(i,j)$, with probability one as $n \to \infty$, we find that $J(i,j)$ is a consistent estimator of $I(i,j)$:

$$J(i,j) \quad \approx \quad \hat{I}(i,j) - \frac{1}{2n}(\alpha^{(i)}-1)(\alpha^{(j)}-1)\log n \qquad (3)$$
$$\to \quad I(i,j) \ ,$$

so that the E minimizing $L(x^n|E)$ converges to the true E as $n \to \infty$ with probability one.

Notice that the resulting $w_{i,j} := J(i,j)$ could be negative, and the resulting graph (V,E) is a forest rather than a tree if we apply the generalized Kruskal algorithm. J. Suzuki (1993) proposed to apply $J(i,j)$ in (3) rather than $\hat{I}(i,j)$ to compare in each step to balance simplicity of the (V,E) and fitness of the x^n to (V,E). On the other hand, the method based on ML estimation considers only fitness to the x^n, and eventually over-fitting occurs although the ML estimator is consistent for large n.

3 Universal Measure and Description Length

3.1 Universal Coding for Finite Sources

Suppose that a sequence of random variables $\{X_i\}_{i=1}^n$ are emitted i.i.d. by probability

[2] $a_n \approx b_n$ denotes $a_n - b_n$ converges to a constant as $n \to \infty$.

$p^n(x^n) = \prod_{i=1}^n p(x_i)$ for $x^n = (x_1, \cdots, x_n)$ with entropy $H(p) := \sum_{x \in A} -p(x) \log p(x)$, where A is a finite set in which each X_i takes values. Then, there exists q^n (see (2)) such that $\sum_{x^n \in A^n} q^n(x^n) \le 1$, and $-\frac{1}{n} \log q^n(x^n) \to H(p)$ for any p with probability one as $n \to \infty$. For example,

$$q^n(x^n) := \frac{\Gamma(\frac{m}{2}) \prod_{a \in A} \Gamma(c_n[a] + \frac{1}{2})}{\Gamma(n + \frac{m}{2}) \Gamma(\frac{1}{2})^m} ,$$

with $m := |A|$ (cardinality of A) satisfies such a property (Cover 1995), where $c_n[a]$ is the number of occurrences of $a \in A$ in $x^n \in A^n$, and Γ is the gamma function. We also notice from the Shannon-McMillan-Breiman theorem that $-\frac{1}{n} \log p^n(x^n) \to H(p)$ for any p, which can be also obtained by the strong law of large numbers:

$$-\frac{1}{n} \log p^n(x^n) = \frac{1}{n} \sum_{i=1}^n - \log p(x_i)$$

($\{- \log p(X_i)\}_{i=1}^n$ are independent random variables). Thus, we have

Proposition 1. There exists q^n such that $\sum_{x^n \in A^n} q^n(x^n) \le 1$, and $\frac{1}{n} \log \frac{p^n(x^n)}{q^n(x^n)} \to 0$ for any p with probability one as $n \to \infty$.

Hereafter, we denote $L(x^n) := - \log q^n(x^n)$.

3.2 Estimation of Density Functions

Suppose that a sequence of random variables $\{X_i\}_{i=1}^n$ are emitted i.i.d. by density function $f^n(x^n) := \prod_{i=1}^n f(x_i)$ for $x^n = (x_1, \cdots, x_n)$. Let $\mathcal{X}$ be a range in which each X_i takes values. We construct a sequence $\{A_k\}_{k=0}^\infty$ such that $A_0 := \{\mathcal{X}\}$ and A_{k+1} is a refinement of A_k.

Example 2. If $\mathcal{X} = [0, 1)$, then $A_0 = \{[0, 1)\}$, and the sequence

$$\begin{array}{rcl} A_1 & = & \{[0, 1/2), [1/2, 1)\} \\ A_2 & = & \{[0, 1/4), [1/4, 1/2), [1/2, 3/4), [3/4, 1)\} \\ & \cdots & \\ A_k & = & \{[0, 2^{-(k-1)}), [2^{-(k-1)}, 2 \cdot 2^{-(k-1)}), \\ & & \cdots, [(2^{k-1} - 1)2^{-(k-1)}, 1)\} \\ & \cdots & \end{array}$$

satisfies the condition.

For each k, we define projection $s_k : \mathcal{X}^n \to A_k^n$ by $x^n \mapsto a^n$ if $x^n \in a^n \in A_k^n$. We denote by λ the Lebesgue measure of $\mathbb{R}$ with $\lambda^n(a^n) = \prod_{i=1}^n \lambda(a_i)$ for $a^n = (a_1, \cdots, a_n)$, and by $p_k^n(a^n) := \prod_{i=1}^n p_k(a_i)$ the probability of $s_k(X^n) = a^n \in A_k^n$.

Since $s_k(X^n)$ is i.i.d., there exists q_k^n (Proposition 1) such that

$$\frac{1}{n} \log \frac{p_k^n(s_k(x^n))}{q_k^n(s_j(x^n))} \to 0$$

for any p_k. If we define $g_k^n(x^n) := \frac{q_k^n(s_k(x^n))}{\lambda^n(s_k(x^n))}$, we construct a measure over $\mathcal{X}^n$ with $\{\omega_k\}_{k=1}^\infty$ such that $\sum \omega_k = 1$, $\omega_k > 0$ to define $g^n(x^n) := \sum_{k=1}^\infty \omega_j g_k^n(x^n)$. Notice $\int_{x^n \in \mathcal{X}^n} g^n(x^n) dx^n = 1$.

Let f_k be the density function associated with A_k, and $h(f)$ the differential entropy of density function f.

Proposition 2 (Ryabko 2009)**.** Fix $\{A_k\}_{j=1}^\infty$. Then, for arbitrary f such that $h(f_k) \to h(f)$ as $j \to \infty$

$$\frac{1}{n} \log \frac{f^n(x^n)}{g^n(x^n)} \to 0 .$$

3.3 Generalization

Lemma 1 (The Radon-Nikodym Theorem (Billingsley, 1995))**.** For σ-finite measures[3] μ, ν on the measure space with entire set Ω and σ-set field $\mathcal{F}$, the following conditions (μ is absolutely continuous with respect to ν) are equivalent:

1. there exists f such that $\mu(D) = \int_D f(x) d\nu(x)$ for $D \in \mathcal{F}$
2. $\mu \ll \nu$, i.e. $\nu(D) = 0 \Longrightarrow \mu(D) = 0$ for $D \in \mathcal{F}$

[3] A measure ν is σ-finite if there exists $\{A_i\}_{i=1}^\infty$ such that $\nu(A_i) < \infty$ and $\cup A_i = \Omega$.

We denote such an f (*Radon-Nikodym derivative* (Billingsley, 1995) of μ with respect to η) by $\dfrac{d\mu}{d\nu}$.

In Section 3.2, we assumed $\mu \ll \lambda$ for an unknown probability measure μ, and construct $g = \dfrac{d\nu}{d\lambda}$ such that $\nu \ll \lambda$.

Now we consider the general case $\mu \not\ll \lambda$. Choose η such that $\mu \ll \eta \neq \lambda$ to estimate the density function $\dfrac{d\mu}{d\eta}$ with respect to η instead. Then, estimation of $\dfrac{d\mu}{d\eta}$ is similar except that λ, λ^n are replaced by η, η^n.

Hereafter, we assume the existence of a underlying supporting σ-finite measure η such that $\mu \ll \eta$, so that $f = \dfrac{d\mu}{d\eta}$ and $g = \dfrac{d\mu}{d\eta}$.

Example 3. Suppose $B = \{1, 2, \cdots\}$, and that the following sequence is given:

$B_0 = \{\{1, 2, \cdots\}\}$
$B_1 = \{\{1\}, \{2, \cdots\}\}$
$B_2 = \{\{1\}, \{2\}, \{3, \cdots\}\}$
$\ldots$
$B_l = \{\{1\}, \cdots, \{l\}, \{l+1, \cdots\}\}$
$\ldots$

For each l, we define projection $t_l : \mathcal{Y}^n \to B^n$ by $y^n \to b^n$ if $y^n \in b^n \in B_l^n$. Then $m = l + 1$, and if the source is i.i.d.,

$$\nu_l^n(t_l(y^n)) := \frac{\Gamma(\frac{m}{2}) \prod_{b \in B_l} \Gamma(c_n[b] + \frac{1}{2})}{\Gamma(n + \frac{m}{2}) \Gamma(\frac{1}{2})^m} \ ,$$

where $c_n[b]$ is the number of occurrences of $b \in B_l$ in $s_l(y^n) \in B_l^n$. If we choose η as $\eta(\{j\}) := \dfrac{1}{j(j+1)}$ for $j = 1, 2, \cdots$, thus

$$\eta(\{l+1, \cdots\}) = \sum_{j=l+1}^{\infty} = \frac{1}{l+1} \ ,$$

then we have $\mu \ll \eta$. Then, we can compute

$$\frac{d\nu_l^n}{d\eta^n} = \frac{\nu_l(t_l(y^n))}{\prod_{i=1}^n \eta(t_l(y_i))}$$

to obtain $\dfrac{d\nu^n}{d\eta^n} = \displaystyle\sum_{l=1}^{\infty} \omega_l \frac{d\nu_l^n}{d\eta^n}$. The obtained measure ν^n is asymptotically close to μ^n in the sense of Theorem 1 as $n \to \infty$.

4 Estimation of Mutual Information

Let X, Y be random variables with ranges $\mathcal{X}$, $\mathcal{Y}$. We assume that the measures μ_X, μ_Y of X, Y are absolutely continuous with respect to σ-finite measures η_X, η_Y, respectively. By $\eta_X \otimes \eta_Y$ we denote the product measure of η_X, η_Y, i.e. $\eta_X \otimes \eta_Y(dx, dy) = \eta_X(dx)\eta_Y(dy)$.

Choose $\{A_k\}, \{B_l\}$ so that the differential entropy of the density functions $f_{X,k}, f_{Y,l}, f_{XY,k,l}$ over A_k, B_l, $A_k \times B_l$ converge to

$$f_X = \frac{d\mu_X}{d\eta_X}, f_Y = \frac{d\mu_Y}{d\eta_Y}, f_{XY} = \frac{d\mu_{XY}}{d(\eta_X \otimes \eta_Y)}$$

as $k, l \to \infty$. Then, there exist g_X^n, g_Y^n, g_{XY}^n such that with probability one as $n \to \infty$ $\dfrac{1}{n} \log \dfrac{f_X^n(x^n)}{g_X^n(x^n)} \to 0$, $\dfrac{1}{n} \log \dfrac{f_Y^n(y^n)}{g_Y^n(y^n)} \to 0$, and $\dfrac{1}{n} \log \dfrac{f_{XY}^n(x^n, y^n)}{g_{XY}^n(x^n, y^n)} \to 0$. Thus,

$$\frac{1}{n} \log \frac{g_{XY}^n(x^n, y^n)}{g_X^n(x^n) g_Y^n(y^n)} - \frac{1}{n} \log \frac{f_{XY}^n(x^n, y^n)}{f_X^n(x^n) f_Y^n(y^n)} \to 0 \ .$$

From the Shannon-McMillan-Breiman theorem,

$$\frac{1}{n} \log \frac{f_{XY}^n(x^n, y^n)}{f_X^n(x^n) f_Y^n(y^n)} = \frac{1}{n} \sum_{i=1}^n \log \frac{f_{XY}(x_i, y_i)}{f_X(x_i) f_Y(y_i)} \to I(X, Y)$$

which can be also obtained from the strong law of large numbers. Thus,

Theorem 1. Given $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$,

$$\frac{1}{n} \log \frac{g_{XY}^n(x^n, y^n)}{g_X^n(x^n) g_Y^n(y^n)}$$

is a consistent estimator of mutual information $I(X, Y)$.

Example 4. We construct $\nu_{k,l}^n(s_k(x^n), t_l(y^n))$ as

$$\frac{\Gamma(\frac{m}{2}) \prod_{a \in A_k} \prod_{b \in B_l} \Gamma(c_n[a, b] + \frac{1}{2})}{\Gamma(n + \frac{m}{2}) \Gamma(\frac{1}{2})^m} \ ,$$

based on the sequence $A_k, B_l,\ k,l = 0,1,2,\cdots$, where $m = 2^k(l+1)$, and $\{A_k\}$ and $\{B_l\}$ have been constructed in Examples 2 and 3, respectively, and $c_n[a,b]$ is the number of occurrences of $a \in A_k$ and $b \in B_l$ in $(s_k(x^n), t_l(y^n)) \in A_k^n \times B_l^n$. Then, we can calculate

$$\frac{d\nu_{k,l}^n}{d\eta^n} = \frac{\nu_{k,l}(s_k(x^n), t_l(y^n))}{\prod_{i=1}^n \lambda(s_k(x_i)) \prod_{i=1}^n \eta(t_l(y_i))}$$

to obtain $\displaystyle\frac{d\nu^n}{d\eta^n} = \sum_{k,l} \omega_{k,l} \frac{d\nu_{k,l}^n}{d\eta^n}$.

5 A Generalized Version of the Chow-Liu Algorithm based on the MDL

Let $g^n(i) := g_X^n(x^n)$ with $X = X^{(i)}$ and a supporting measure η_i, and $g^n(i,j) := g_{XY}^n(x^n, y^n)$ with $X = X^{(i)}$, $Y = X^{(j)}$ and the supporting measure $\eta_i \otimes \eta_j$. We define the Bayesian measure over $\mathcal{X}^n$ relative to (V,E), where $\mathcal{X} := \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(N)}$

$$g^n(1,\cdots,N|E) := \frac{\prod_{\{i,j\}\in E} g^n(i,j)}{\prod_{i\in V} g^n(i)^{d_i - 1}} = \prod_{\{i,j\}\in E} \frac{g^n(i,j)}{g^n(i)g^n(j)} \prod_{i \in V} g^n(i) \ ,$$

the description length with respect to (V,E) and $\eta := \eta_1 \otimes \cdots \otimes \eta_N$

$$L_\eta(x^n|E) := -\log g^n(1,\cdots,N|E) = -\sum_{i\in V} \log g^n(i) - \sum_{\{i,j\}\in E} \log \frac{g^n(i,j)}{g^n(i)g^n(j)} \ ,$$

and the Bayesian estimator

$$J(i,j) \ := \ \frac{1}{n} \log \frac{g^n(i,j)}{g^n(i)g^n(j)} \ .$$

From Theorem 1, we can choose as $J(i,j)$ a consistent estimator of $I(i,j)$.

Example 5. Let $\{A_k\}, \{B_l\}, \{A_k \times B_l\}$ be the sequences constructed in Examples 2,3 and 4. In order to obtain the score $g^n(i)$, we consider $(\{A_k\}, \lambda)$. Given $x^n = (x_1, \cdots, x_n) \in A^n$, we obtain the score $g^n(i)$ for the weights $\{w_k\}_{k=1}^K$ and functions s and λ. For $k = 1, \cdots, K$, the s returns $a \in A_k$ such that $x_j \in a$ for each x_j, and the λ returns the measure of $a \in A_k$, where $|A_k|$ denotes the cardinality of A_k.

For each $k = 0,1,\cdots,K$, the following algorithm returns $g_k^n(x^n)$ given x^n and A_k, and we obtain $g^n(i) := \sum_l w_k g_k^n(x^n)$ given $\{w_k\}$.

```
1. $c[a] := 0$ for $a \in A_k$;
2. $g_k^n := 1$;
3. for $h = 1, \cdots, n$
    (a) $a := s(x_j)$;
    (b) $c[a] := c[a] + 1$;
    (c) $g_k^n := g_k^n * \frac{c[a] + 1/2}{h + |A_k|/2} / \lambda(a)$;
```

It is easy to find that the computation in the proposed algorithm is linear with number n of the examples. So, if the computation is large, then that will be due to the choice of $\{A_k\}_{k=1}^K$.

If we apply binary search for step 3 (a), then $\log_2 |A_k|$ comparisons are required. So, at most $O(L \log |A_k|)$ computation is required for the K cycles. Some might think that $|A_K|$ would be eventually large. But in reality, we cannot make $|A_k|$ so large unless n is fairly large. In fact, if $|A_k|$ is too large, then each $c[a]$ will be small, so that the value of g_k^n is not significant compared with $\{g_r^n\}_{r<k}$, and does not affect the resulting value of so much.

Similarly, for each $k = 0,1,\cdots,K$ and $l = 0,1,\cdots,L$ the following algorithm returns $g_{kl}^n(x^n)$ given x^n and A_k, B_l, and we obtain $g^n(i,j) := \sum_{k,l} w_{kl} g_{kl}^n(x^n)$ given $\{w_{kl}\}$.

```
1. $c[a,b] := 0$ for $a \in A_k, b \in B_l$;
2. $g_{kl}^n := 1$;
3. for $h = 1, \cdots, n$
    (a) $a := s(x_j)$; $b := t(y_j)$;
    (b) $c[a,b] := c[a,b] + 1$;
    (c) $g_{kl}^n := g_{kl}^n * \frac{c[a,b] + 1/2}{h + |A_k||B_l|/2} / (\lambda(a)\eta(b))$;
```

Table 1 shows the values $J(i,j)$ and its computation time for $I(i,j) = 2.0$ (The logarithm base is two) and $K = 8, 64$ and $n = 100, 1000$.

(K, L)	n	$J(i,j)$	time (ms)
(2,4)	100	0.0612	1.23
(2,4)	1000	0.0531	10.67
(4,8)	100	0.0428	1.69
(4,8)	1000	0.0342	14.71

Table 1: The value $J(i,j)$ and its averaged computation time

We generate $(x^n, y^n) \in (A \times B)^n$ one hundred times to obtain the arithmetic average of $J(i,j)$. The data (x^n, y^n) are generated so that they are independent $(I(X,Y) = 0)$. We find that if K is too small, we only obtain an approximation even when n is large, and that for large K, large n is required for convergence.

In the current problem, for any method, the computation is eventually high, but this is due to the nature of the problem, not due to the proposed method.

6 Concluding Remarks

In this paper, we proposed the Bayesian estimator of mutual information which has several merits:

1. consistent: the true mutual information is obtained as n grows;
2. Bayesian: maximizing the sum of mutual information over the chosen edges leads to minimizing the description length relative to the forest for each n;
3. nonparametric: no assumption about any specific parameters is required;
4. robust: compared to existing approaches (Wang et al, 2005)(Silva and Narayan 2010) because the proposed approch does not seek any one histogram but evaluates mutual information based on mixtured values of those candidate histograms; and
5. general: applicable to any random variable.

Future works includes figuring out more applications of the universal Bayesian measure. Thus far, it has been found that a similar approach can be applied to Bayesian network structure estimation when both discrete and continuous variables are present (Suzuki 2011).

References

P. Billingsley. *Probability & Measure* (1995): (3rd ed.). New York : Wiley.

C. K. Chow and C. N. Liu. “Approximating discrete probability distributions with dependence trees”. *IEEE Transactions on Information Theory*, IT-14(3):462–467, May (1968).

T. M. Cover and J. A. Thomas. *Elements of Information Theory* (1995): (2nd ed.). New York : Wiley. *Elements of information theory*: John Wiley & Sons, New York, NY, (1991).

G. A. Darbellay and I. Vajda. “Estimation of the information by an adaptive partitioning of the observation space”. *IEEE Trans. Information Theory*, 45(4):1315-1321, 5 (1999).

R.E. Krichevsky and V.K. Trofimov, ’The Performance of Universal Encoding”, *IEEE Trans. Information Theory*, Vol. IT-27, No. 2, pp. 199-207 (1981)

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufmann (1988)

J.Rissanen, “Modeling by shortest data description”. *Automatica* 14: 465-471 (1978).

B. Ryabko. “Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series.” *IEEE Trans. on Inform. Theory*, 55(9):4309-4315 (2009).

Jorge Silva and Shrikanth Narayanan, Non-product data-dependent partitions for mutual information estimation: Strong consistency and applications, *IEEE Transactions on Signal Processing* (2010).

J. Suzuki “A Construction of Bayesian Networks from Databases on the MDL principle”, *Uncertainty in Artificial Intelligence*, Washington DC, July 1993.

J. Suzuki, “The Universal Measure for General Sources and its Application to MDL/Bayesian Criteria”, Data Compression Conference 2011, Snowbird, Utah, 2011, page 478 (one page abstract).

Q. Wang, S. Kulkarni, and S. Verd’u. “Divergence estimation of continuous distributions based on data-dependent partitions.” *IEEE Trans. Information Theory*, 51(9):3064-3074, 9 (2005).