2005－MPS－56（19）
2005／9／22

ベイジアンネットワークの構造学習における強一致性について

鈴木譲
Joe Suzuki*1

大阪大学 大学院理学研究科
Graduate School of Science, Osaka University

Abstract: 帰納推論でも、統計学でも、データマイニングでも、トレーニング例の数 n が増えてくれば、正しい学習結果が欲しくなる (一致性)。情報量基準でベイジアンネットワークの構造学習を行う際に、ネットワークの複雑さに対するペナルティを n の関数 $c(n)$ にパラメータ数を乗じた値であらわすときに、$c(n)$ が小さすぎると、過学習となり、一致性がいえなくなる。一致性が成立する最小の $c(n)$ を、一般的に求めた。

(The proofs of Theorems 1 and 2 are abbreviated because of space constraint.)

1 Introduction

Suppose we are learning a Bayesian network (BN) structure from finite examples (Cooper and Herskovits, 1992).

A BN is a directed acyclic graph with some nodes $1, 2, \cdots, N$ (N: the number of nodes) and edges directing from each $k \in \pi^{(j)}$ to j, for some $\pi^{(j)} \subseteq \{1, 2, \cdots, j-1\}$, $j = 1, \cdots, N$. By the BN structure, we mean the $\pi^N = (\pi^{(1)}, \cdots, \pi^{(N)})$ of the BN. Each node j corresponds to a random variable $X^{(j)}$. We assume $X^{(j)}$ takes on a finite set $\mathcal{X}^{(j)}$ with cardinality $\alpha^{(j)}$, and that the marginal distribution of $X^{(1)}, \cdots, X^{(N)}$ is expressed by

$$\begin{aligned} & P(X^{(1)} = x^{(1)}, \cdots, X^{(N)} = x^{(N)}) \\ = & \prod_{j=1}^{N} P(X^{(j)} = x^{(j)} | \{X^{(k)} = x^{(k)}\}_{k \in \pi^{(j)}}) \quad (1) \end{aligned}$$

$x^{(1)} \in \mathcal{X}^{(1)}, \cdots, x^{(N)} \in \mathcal{X}^{(N)}$, i.e. the occurrence of $X^{(j)}$ depends on those of $X^{(k)}$, $k \in \pi^{(j)}$. If the Bayesian network structure π^N is given, the marginal distribution is specified if for $j = 1, \cdots, N$, $P(X^{(j)} = x^{(j)} | X^{(k)} = x^{(k)}, k \in \pi^{(j)})$ are given for all $x^{(j)} \in \mathcal{X}^{(i)}, x^{(k)} \in \mathcal{X}^{(k)}, k \in \pi^{(j)}$.

The problem is to estimate $\pi = (\pi^{(1)}, \cdots, \pi^{(N)})$ from n examples x^n

$$\begin{array}{llll} X^{(1)} = x_1^{(1)}, & X^{(2)} = x_1^{(2)}, & \cdots, & X^{(N)} = x_1^{(N)} \\ X^{(1)} = x_2^{(1)}, & X^{(2)} = x_2^{(2)}, & \cdots, & X^{(N)} = x_2^{(N)} \\ \cdots, & \cdots, & \cdots, & \cdots \\ X^{(1)} = x_n^{(1)}, & X^{(2)} = x_n^{(2)}, & \cdots, & X^{(N)} = x_n^{(N)} , \end{array}$$

assuming that the x^n have been emitted independently and identically distributed according to (1) with the structure π^N, and that no value is missing in the nN attributes. Hence, without loss of generality, we estimate each $\pi^{(j)}$ independently, $j = 1, 2, \cdots, N$.

We define $\mathcal{S}(\pi^{(j)}) := \{\{(x^{(1)}, \cdots, x^{(j-1)}) | x^{(k)} = s^{(k)}, k \in \pi^{(j)}\} | s^{(k)} \in \mathcal{X}^{(k)}, k \in \pi^{(j)}\}$. Then, $\mathcal{S}(\pi^{(j)})$ has $\prod_{k \in \pi^{(j)}} \alpha^{(k)}$ elements, and for each $s = \{(x^{(1)}, \cdots, x^{(j-1)}) | x^{(k)} = s^{(k)}, k \in \pi^{(j)}\} \in \mathcal{S}(\pi^{(j)})$, we notice

$$\begin{aligned} & P(X^{(j)} = x^{(j)} | X^{(k)} = s^{(k)}, k \in \pi^{(j)}) \\ = & P(X^{(j)} = x^{(j)} | (X^{(1)}, \cdots, X^{(j-1)}) \in s) . \end{aligned}$$

Hereafter, let $\mathcal{X} := \mathcal{X}^{(1)} \times \mathcal{X}^{(j-1)}$, $\mathcal{Y} := \mathcal{X}^{(j)}$, $X := (X^{(1)}, \cdots, X^{(j-1)})$, $Y := X^{(j)}$, and $\pi := \pi^{(j)}$. Let $\pi_* = \pi_*^{(j)}$ be the true $\pi = \pi^{(j)}$, and $p[y, s^*] := P(Y = y | X \in s^*)$ for each $s^* \in \mathcal{S}(\pi^*)$. We define $p[s] := P(X \in s)$ for each $s \subseteq \mathcal{X}$, and

$$p[y, s] := \sum_{s^* \in \mathcal{S}(\pi^*)} \frac{p[s \cap s^*]}{p[s]} p[y, s^*] . \quad (2)$$

for $\pi \neq \pi^*$ and $s \in S(\pi)$ such that $p[s] > 0$. Notice $p[y, s]$ is the sum of $\{p[y, s^*]\}_{s^* \in \mathcal{S}(\pi_*)}$ weighted by $p[s \cap s^*]/p[s]$, and that for $s \in \pi$ and $s^* \in \pi^*$

$$s \subseteq s^* \Longrightarrow p[y, s] = p[y, s^*] . \quad (3)$$

鈴木譲, 大阪大学大学院理学研究科数学専攻
573-1194 豊中市待兼山町 1-1
E-Mail: suzuki@math.sci.osaka-u.ac.jp

For each $y \in \mathcal{Y}$ and $s \in S(\pi)$, let $c_n[y,s]$ is the number of occurrences in x^n such that $Y = y$ and $X \in s$, and $c_n[s] := \max\{1, \sum_{y \in \mathcal{X}^{(j)}} c_n[y,s]\}$. Then, $\hat{p}_n[y,s](x^n) := c_n[y,s]/c_n[s]$ almost surely converges to $p[y,s]$.

In this paper, we analyze the following strategy: select the model minimizing

$$L(\pi, x^n) := H(\pi, x^n) + \frac{k(\pi)}{2} d_n \tag{4}$$

(Suzuki, 1993), where $H(\pi, x^n) := \sum_{s \in \mathcal{S}(\pi)} \sum_{y \in \mathcal{Y}} c_n[y,s] \log \frac{c_n[s]}{c_n[y,s]}$, $k(\pi) := (|\mathcal{Y}| - 1)|\mathcal{S}(\pi)| = (\alpha^{(j)} - 1) \prod_{k \in \pi} \alpha^{(k)}$, $\{d_n\}_{n=1}^{\infty}$ is a real nonnegative sequence, $\log x$ denotes the natural logarithm of x, and $0 \log 0 = 0$.

$H(\pi, x^n)$ is the so-called empirical entropy of $y \in \mathcal{Y}$ given $s \in \mathcal{S}(\pi)$, and $k(\pi)$ is interpreted as the number of independent $\{p[y,s]\}_{y \in \mathcal{Y}, s \in \mathcal{S}(\pi)}$ because for each $s \in \mathcal{S}(\pi)$, $\sum_{y \in \mathcal{Y}} p[y,s] = 1$.

We compromize fitness of data to a model and simplicity of the model by balancing $H(\pi, x^n)$ and $k(\pi)$ and adjust their weights by $\{d_n\}_{n=1}^{\infty}$.

This paper analyzes the model selection procedures based on information criteria in the form of (4) in a unified manner, instead of considering each information criterion such as Akaike's information criterion (Akaike 1974, $d_n = 2$), the minimum description length (MDL) principle (Rissanen 1978, $d_n = \log n$), the Bayesian information criterion (Schwarz 1978, $d_n = \log n$), etc.

For each $j = 1, 2, \cdots, N$, we classify the estimation result $\hat{\pi}_n(x^n) := \mathrm{argmin}_{\pi} L(\pi, x^n)$ by $\hat{\pi}_n(x^n) = \pi_*$ (the exact links), $\hat{\pi}_n(x^n) \supset \pi_*$ (adding extra links without missing links), and $\hat{\pi}_n(x^n) \not\supseteq \pi_*$ (missing links). We wish the probability of the first category to be as large as possible for each j.

We define the Kullback divergence by

$$D(\pi_* || \pi) := \sum_{s \in \mathcal{S}(\pi)} \sum_{s^* \in \mathcal{S}(\pi_*)} \sum_{y \in \mathcal{Y}} p[s \cap s^*] p[y, s^*] \log \frac{p[y, s^*]}{p[y,s]} .$$

Then, if $\pi \supseteq \pi_*$, for all $\forall s \in \mathcal{S}(\pi)$, there exists $s^* \in \pi_*$ such that $s \subseteq s^*$, so that $D(\pi_* || \pi) = 0$ (see (3)).

$$\pi \supseteq \pi_* \Longrightarrow D(\pi_* || \pi) = 0 . \tag{5}$$

Let $s(x, \pi), x \in \mathcal{X}$ be the $s \in \mathcal{S}(\pi)$ such that $x \in s$, and

$$s(\pi_*, \pi) := \{x \in \mathcal{X} | p[y, s(x,\pi)] \neq p[y, s(x, \pi_*)]\}$$

be the set of $x \in \mathcal{X}$ such that the probabilities of $y \in \mathcal{Y}$ are different between the structures π_* and π. We assume that $P(X = x)$ satisfies

Assumption 1 For π such that $s(\pi_*, \pi) \neq \phi$, $P(X \in s(\pi_*, \pi)) > 0$,

which means that we can obtain examples that distinguish π_* and π such that $s(\pi_*, \pi) \neq \phi$. Thus,

$$D(\pi_* || \pi) = 0 \Longrightarrow \pi \supseteq \pi_* . \tag{6}$$

We derive (in Section 2) the asymptotic exact error probability $P(\hat{\pi}_n(x^n) \neq \pi_*)$ in model selection for each $\{d_n\}_{n=1}^{\infty}$ in (4). Then, if $\hat{p}_n[y,s](x^n)$ almost surely converges to $p[y,s]$, from (5), the discriminant betwen π_* and π ($\not\supseteq \pi_*$) is asymptotically larger than that between π_* and π ($\supset \pi_*$). We show in Section 2 that the probability of $\hat{\pi}(X^n) \neq \pi_*$ is expressed in terms of the $k(\pi) - k(\pi_*)$ and $\{d_n\}_{n=1}^{\infty}$ for the second category while that of $\hat{\pi}_n(X^n) \neq \pi_*$ exponentially diminishes for the third category:

$$P\{x^n : \hat{\pi}_n(x^n) \neq \pi_*\} \leq 1 - \frac{\Gamma_{\frac{k(\pi)-k(\pi_*)}{2} d_n}(\frac{k(\pi) - k(\pi_*)}{2})}{\Gamma(\frac{k(\pi) - k(\pi_*)}{2})} ,$$

where $\Gamma_x(\cdot)$ is the incomplete Gamma function with respect to parameter x.

There are two meanings of consistency in model selection:

- if $P(\hat{\pi}_n(X^n) \neq \pi_*) \to 0$ as $n \to \infty$ (probability convergence), then we say the sequence $\{d_n\}_{n=1}^{\infty}$ is weekly consitent; and
- if $X^{\infty} = x^{\infty}$ such that $\hat{\pi}(X^n) = \pi_*$ for all but finite n with probability one (almost sure convergence), we say the sequence $\{d_n\}_{n=1}^{\infty}$ is strongly consistent.

Strong consistency implies weak consistency although the converse is not true. We are interested in strong consistency.

Let D be the set of real nonnegative sequences $\{d_n\}_{=1}^{\infty}$ such that $\limsup_{n \to \infty} d_n/n = 0$. We define the partial order $<$ in D: for any $\{d\}_{n=1}^{\infty}, \{d'\}_{n=1}^{\infty} \in D$, $\{d\}_{n=1}^{\infty} > \{d'\}_{n=1}^{\infty}$ if $\liminf_{n \to \infty} \frac{d_n}{d'_n} > 1$.

The climax of this paper (in Section 3) is in the derivation of the smallest $\{d\}_{n=1}^{\infty} \in D$ that satisfies strong consistency for BN structure learning,

i.e. the BN structure learning counterpart of Hannan and Quinn's information criterion that was provided for ARMA processes. More precisely, the problem is whether $d_n^* = 2\log\log n$ is the smallest such that $\{d_n\}_{n=1}^\infty$ is strongly consistent for $\forall \pi_*^N = (\pi_*^{(1)}, \cdots, \pi_*^{(N)})$. We solve this problem in the affirmative.

2 Error probabilities

In what follows, we derive $P\{x^n : \hat{\pi}_n(x^n) \neq \pi_*\}$ both for the two cases: $\hat{\pi}_n(x^n) \supset \pi_*$ and $\hat{\pi}_n(x^n) \not\supseteq \pi_*$.

2.1 Error for adding extra links without missing links

Theorem 1 *For $\hat{\pi}_n(x^n) \supset \pi_*$ and any $\{d_n\}_{n=1}^\infty \in D$.*

$$P\{x^n : \hat{\pi}_n(X^n) \neq \pi_*\} \leq 1 - \frac{\Gamma_{\frac{k(\pi)-k(\pi_*)}{2} d_n}(\frac{k(\pi)-k(\pi_*)}{2})}{\Gamma(\frac{k(\pi)-k(\pi_*)}{2})} \tag{7}$$

almost surely as $n \to \infty$, where $\Gamma_x(\cdot)$ is the incomplete Gamma function with respect to parameter x and $\Gamma(\cdot) = \Gamma_\infty(\cdot)$ is the Gamma function.

2.2 Error for missing links

Theorem 2 *For $\hat{\pi}_n(x^n) \not\supseteq \pi_*$, almost surely as $n \to \infty$*

$$\frac{L(\pi, x^n) - L(\pi_*, x^n)}{n} \to D(\pi_* || \pi) \ . \tag{8}$$

In particular, almost surely as $n \to \infty$, $\hat{\pi}_n(x^n) \neq \pi_*$ for any $\{d_n\}_{n=1}^\infty \in D$.

3 Strong Consistency of Model Selection

Theorem 3 suggests that $d_n^* = 2\log\log n$ is the smallest in D that makes the model selection procedure strongly consistent for BN structure learning as well as for ARMA processes.

Theorem 3 Under Assumption 1,

1. If $\{d^*\}_{n=1}^\infty < \{d\}_{n=1}^\infty \in D$, then $\{d_n\}_{n=1}^\infty$ is strongly consistent for $\forall \pi_*^N$; and
2. If $\{d^*\}_{n=1}^\infty > \{d\}_{n=1}^\infty \in D$, then $\{d_n\}_{n=1}^\infty$ is not strongly consistent for $\exists \pi_*^N$ if $N \geq 2$,

where $\pi_*^N = (\pi_*^{(1)}, \cdots, \pi_*^{(N)})$.

Proof of Theorem 3: From Theorem 2, for $\pi \not\supseteq \pi_*$, $L(\pi, x^n) > L(\pi_*, x^n)$ almost surely as $n \to \infty$ if $\{d_n\}_{n=1}^\infty \in D$.

We show for π ($\supset \pi_*$), almost surely

$$\limsup_{n\to\infty} \frac{H(\pi_*, x^n) - H(\pi, x^n)}{\log\log n} = K(\pi, \pi_*) \tag{9}$$

where $K(\pi, \pi_*) = \sum_{s^* \in \mathcal{T}} (|\mathcal{T}(s^*, \pi)| - 1)(|\mathcal{Y}| - 1)$ as before.

The proof is based on an extended version of Kolmogorov's law of the iterated logarithm. (Stout 1974, page 269 for independent and identically distributed random variables):

Lemma 1 *Let $\{S_n, \mathcal{F}_n, n \geq 1\}$ be a martingale with $S_0 = 0$ and $E[S_1] = 0$. Let K_n be $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$ with $K_n \to 0$ a.s., and $Z_n = S_n - S_{n-1}$ for $n \geq 1$. If we suppose $s_n^2 = \sum_{t=1}^n E[Z_t^2 | \mathcal{F}_{t-1}] \to \infty$ as $n \to \infty$ and*

$$Z_n \leq \frac{K_n s_n}{\sqrt{\log\log s_n^2}} \ a.s. \tag{10}$$

for $n \geq 1$, then almost surely

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2 s_n^2 \log\log s_n^2}} = 1 \tag{11}$$

We define $\{Z_n\}$ as

$$\sum_{s \in S(s^*, \pi_*)} \sum_{y \in \mathcal{Y}} u[i,s] w[y,j] \frac{I[X_n \in s](I[Y_n = y] - p[y, s^*])}{\sqrt{c_n[s] p[y, s^*]/n}} \ . \tag{12}$$

One easily checks that $\{S_n, \mathcal{F}_n, n \geq 1\}$ is a martingale, i.e. (12) satisfies $E[Z_n | \mathcal{F}_{n-1}] = 0$ with $Z_n = S_n - S_{n-1}$ for each $n \geq 1$, $S_0 = 0$, and $E[S_1] = 0$. Then, from the definitions of matrices V, U, W, and Q,

$$\begin{aligned} S_n &= \sum_{t=1}^n Z_t \\ &= \sum_{s \in S(s^*, \pi)} \sum_{y \in \mathcal{Y}} u[i,s] w[y,j] \{ \sum_{t=1}^n I[X_t \in s, Y_t = y] \\ &\quad - \sum_{t=1}^n I[X_t \in s] p[y, s^*] \} / \sqrt{c_n[s] p[y, s^*]/n} \\ &= \sqrt{n} \sum_{s \in S(s^*, \pi_*)} \sum_{y \in \mathcal{Y}} u[i,s] w[y,j] v_n[y,s] = \sqrt{n} q[i,y] \ . \end{aligned}$$

Since

$$E[I[X_t \in s](I[Y_t = y] - p[y, s^*])I[X_{t'} \in s'](I[Y_{t'} = y'] - p[y', s^*])/\sqrt{c_n[s]c_n[s']}] = 0$$

for $s, s' \in S(s^*, g)$, $y, y' \in \mathcal{Y}$, where $E[\cdot]$ is the expectation over (X_t, Y_t), $t = 1, 2, \cdots, n$, we have $E[Z_t Z_{t'}] = 0$ for $1 \leq t < t'$, so that

$$\begin{aligned} s_n^2 &= \sum_{t=1}^{n} E[Z_t^2 | \mathcal{F}_{t-1}] \\ &= E[\{\sum_{t=1}^{n} Z_t\}^2] s_n^2 = E[S_n^2] \\ &= E[\{\sqrt{n} q[i,j]\}^2] = n \ . \end{aligned} \tag{13}$$

If we put $K_n = n^{-1/3}$, the conditions for Lemma 3 are satisfied: $K_n \to 0$ as $n \to \infty$; $s_n^2 = n \to \infty$ as $n \to \infty$; and $|Z_n| < \infty$ and $K_n s_n / \sqrt{2 \log \log s_n^2} = n^{1/6} / \sqrt{2 \log \log n} \to \infty$ as $n \to \infty$. Therefore, from (12) and (13), we obtain

$$\limsup_{n \to \infty} \frac{\sqrt{n} q[i,j]}{\sqrt{2n \log \log n}} = 1$$

a.s., $i = 1, 2, \cdots, \alpha - 1$, $j = 1, 2, \cdots, \beta - 1$.

Then, from Proposition 1, almost surely,

$$\begin{aligned} & \limsup_{n \to \infty} \frac{\mathcal{A}(s^*, \pi, x^n)}{2 \log \log n} \\ = & \limsup_{n \to \infty} \frac{\sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} q[i,j]^2}{2 \log \log n} \\ = & \sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} \limsup_{n \to \infty} \frac{q[i,j]^2}{2 \log \log n} \\ = & \sum_{i=1}^{\alpha-1} \sum_{j=1}^{\beta-1} 1 = (\alpha - 1)(\beta - 1) \\ = & (|\mathcal{T}(s^*, \pi)| - 1)(|\mathcal{Y}| - 1) \ . \end{aligned} \tag{14}$$

The second equality in (14) should be generally "$\leq$". However, from Theorem 1, $q[i,j]$, $i = 1, 2, \cdots, \alpha - 1, j = 1, 2, \cdots, \beta - 1$, are statistically independent. Therefore, each term $q[i,j]$ can independently reach the limsup infinitely many times but exceed it finitely many times with probability one. Therefore, the second equality in (14) holds. From (**??**), (9) follows.

If $\{2 \log \log n\}_{n=1}^{\infty} < \{d_n\}_{n=1}^{\infty} \in D$, then almost surely

$$\begin{aligned} & \liminf_{n \to \infty} \frac{L(\pi, x^n) - L(\pi_*, x^n)}{2 \log \log n} \\ = & - \liminf_{n \to \infty} \frac{H(\pi_*, x^n) - H(\pi, x^n)}{2 \log \log n} \\ & + \frac{k(\pi) - k(\pi_*)}{2} \frac{d_n}{2 \log \log n} \\ \geq & \frac{1}{2}(k(\pi) - k(\pi_*))(\frac{d_n}{2 \log \log n} - 1) \ . \end{aligned} \tag{15}$$

Since $\pi \neq \pi_*$, $k(\pi_*) < k(\pi)$, so that the right of (15) is strictly positive for large n. Hence, almost surely $\hat{\pi}_n(x^n) = \pi_*$.

On the other hand, suppose $\{2 \log \log n\} > \{d_n\}_{n=1}^{\infty} \in D$. Then, for $G = \{\pi_*, \pi\}$ such that $\pi_*^N = (\{\}, \cdots, \{\})$ and $\pi^N = (\{\}, \{1\}, \cdots, \{\})$ $(N \geq 2)$, almost surely

$$\begin{aligned} & \liminf_{n \to \infty} \frac{L(\pi, x^n) - L(\pi_*, x^n)}{2 \log \log n} \\ = & \frac{|\mathcal{Y}|(|\mathcal{X}^{(1)}| - 1)}{2} (\frac{d_n}{2 \log \log n} - 1) < 0 \ . \end{aligned}$$

Hence, $L(\pi, x^n) < L(\pi_*, x^n)$ infinitely many times with probability one. This completes the proof.

References

H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716-723. 1974

A. C. Atkinson, "A Method for discriminating between Models", *J. Roy. Statist., Soc. Ser.*, Vol. B32, pp. 323-353, 1970.

P. Billingsley, *Statistical inference for Markov processes*, The University of Chicago Press, 1961.

H. Cramer, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1946.

Cooper, G. F. and E. Herskovits (1992). "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning* **9**: 309-347.

E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression", *J. Roy. Statist. Soc.*, Ser. B. 41, pp. 190-195. 1979.

J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, pp. 465-471. 1978;

G. Schwarz, "Estimating the dimension of a model", *Annals Statist.*, Vol. 6, pp. 461-464, 1978.

W. F. Stout, *Almost Sure Convergence*, Academic Press, 1974.

J. Suzuki. "A Construction of Bayesian Networks from Databases Based on the MDL principle", *the 1993 Uncertainty in Artificial Intelligence conference*: 266-273 (1993); "Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties" *IEICE Trans. on Foundamentals* **E82-A**: 2237-2245 (1999).