Journal of Machine Learning Research 10 (2009) 1511-1526 Submitted 2/08; Revised 3/09; Published 7/09

Strong Limit Theorems for the Bayesian Scoring Criterion in Bayesian Networks

Nikolai Slobodianik NIKOLAI.SLOBODIANIK@GMAIL.COM
Department of Mathematics and Statistics
York University
4700 Keele Street, Toronto, ON M3J 1P3, Canada

Dmitry Zaporozhets ZAP1979@GMAIL.COM
Steklov Institute of Mathematics at St. Petersburg
27 Fontanka, St. Petersburg 191023, Russia

Neal Madras MADRAS@MATHSTAT.YORKU.CA
Department of Mathematics and Statistics
York University
4700 Keele Street, Toronto, ON M3J 1P3, Canada

Editor: Max Chickering

Abstract

In the machine learning community, the Bayesian scoring criterion is widely used for model selection problems. One of the fundamental theoretical properties justifying the usage of the Bayesian scoring criterion is its consistency. In this paper we refine this property for the case of binomial Bayesian network models. As a by-product of our derivations we establish strong consistency and obtain the law of iterated logarithm for the Bayesian scoring criterion.

Keywords: Bayesian networks, consistency, scoring criterion, model selection, BIC

1. Introduction

Bayesian networks are graphical structures which characterize probabilistic relationships among variables of interest and serve as a ground model for doing probabilistic inference in large systems of interdependent components. A basic element of a Bayesian network is a directed acyclic graph (DAG) which is bound to an underlying joint probability distribution by the Markov condition. The absence of certain arcs (edges) in a DAG encodes conditional independences in this distribution. DAG's not only provide a starting point for implementation of inference and parameter learning algorithms, but they also, due to their graphical nature, offer an intuitive picture of the relationships among the variables. It happens too often that researchers have only a random sample from a probability distribution and face a problem of choosing the appropriate DAG between a large number of competing structures. This, effectively, constitutes the model selection problem in the space of Bayesian networks. The methodology which is concerned with solving such task is called Bayesian structure learning.

Suppose that the data consists of n i.i.d. random vectors $X_1,\dots,X_n$ with each X_i having the unknown probability distribution P. We define a probability space Ω with measure Pr for infinite i.i.d. sequences $X_1,X_2,\dots$ having distribution P. There are many structures which can form a Bayesian

network with the distribution P (see Section 2 for formal definitions and examples), however not all of them are optimal for future analysis. Indeed, since the presence of an arc (edge) does not necessarily guarantee direct dependency between corresponding variables, a complete DAG constitutes a Bayesian network with any probability distribution, yet provides no information about conditional independences in P. It is natural to seek structures which not only form a Bayesian network with P, but also entail only conditional independences in this distribution. These DAGs are called *faithful* to P or else *perfect maps* of P. Unfortunately, it turns out that not all probability distributions have an associated faithful structure. In this case it is desirable to identify a structure which satisfies certain "optimality" properties with respect to P. Roughly speaking, we want to include only those edges that are necessary for describing P.

A *scoring criterion* for DAGs is a function that assigns a value to each DAG under consideration based on the data. Suppose M is the set of all DAGs of a fixed size. Under the Bayesian approach to structure learning, the DAG m is chosen from M such that m maximizes the posterior probability given the observed data D:

$$p(m|D,\psi) = \frac{p(m|\psi)p(D|m,\psi)}{p(D|\psi)} = \frac{p(m|\psi)\int_{\Omega_m} p(D|m,\Theta_m,\psi)p(\Theta_m|m,\psi)d\Theta_m}{\sum_{m\in M} p(m|\psi)\int_{\Omega_m} p(D|m,\Theta_m,\psi)p(\Theta_m|m,\psi)d\Theta_m}, \qquad (1)$$

where Θ_m denotes the set of parameters of the conditional distributions of each "node given its parents" for all the nodes of the DAG m, Ω_m denotes the domain of these parameters, and ψ denotes the system of parameter priors. The quantity $P(D|m,\psi)$ is called the *marginal likelihood*, *Bayesian scoring criterion* or else *Score* of the graph m. We denote it as $\mathtt{score}_B(D|m)$. Assuming $\sum_{m\in M} p(m|\psi) = 1$ for all $m \in M$, the Bayesian scoring criterion provides a measure of posterior certainty of the graph m under the prior system ψ.

It is quite interesting to see if the Bayesian scoring criterion is *consistent*, that is, as the size of data D approaches infinity, the criterion is maximized at the DAG which forms a Bayesian network with P and has smallest dimension. Based on the fundamental results of Haughton (1988) and Geiger et al. (2001), the consistency of Bayesian scoring criterion has been established for the class of multinomial Bayesian networks. Chickering (2002) provides a detailed sketch of the proof. Further, for the same model class, if P admits a faithful DAG representation m, then m has the smallest dimension among all DAGs which form a Bayesian network with P (see, for example, Neapolitan, 2004, Corollary 8.1) . Therefore, due to consistency of the Bayesian scoring criterion, we can conclude that if P admits a faithful DAG representation m then, in the limit, the Bayesian scoring criterion will be maximized at m. This last result is naturally expected: as more information becomes available, a scoring criterion should recognize the properties of the underlying distribution P with increasing precision.

Although the consistency property provides insight into the limiting properties of the posterior distribution over the graph space, it is interesting to know at what rate (as a function of sample size) the graph(s) with the smallest dimension become favored by the Bayesian scoring criterion. In this article we address this question for the case of binomial Bayesian network models. We also show that in addition to being consistent for these models, the Bayesian scoring criterion is also *strongly consistent* (see Definition 4). Our proofs are mostly self-contained, relying mainly on well-known limit theorems of classical probability. At one point we require the input of Haughton (1988) and Geiger et al. (2001) mentioned in the preceding paragraph (but note that their results only deal with consistency, not strong consistency).

It may be possible to re-derive our results using the machinery of VC classes (Vapnik, 1998) or empirical process theory (e.g., van der Vaart and Wellner, 1996), but to our knowledge this has not yet been done. However, one point of our paper is to show that the results are amenable to fairly transparent and accessible proofs, and do not require the overhead of these well-developed theoretical frameworks. That being said, we note that our method assumes that the networks have fixed finite size, and other approaches may be better suited to handling the situation in which the network size gets large.

The rest of the paper is organized as follows. Background and notation appear in Section 2, with some illustrative examples. Our results are presented in Section 3. Section 4 contains some discussion. Proofs appear in the Appendix.

2. Background

A directed graph is a pair (V,E), where $V = \{1,\ldots,N\}$ is a finite set whose elements are called nodes (or vertices), and E is a set of ordered pairs of distinct components of V. Elements of E are called edges (or arcs). If $(i_1,i_2) \in E$ we say that there is an edge from i_1 to i_2. Given a set of nodes $\{i_1,i_2,\ldots,i_k\}$ where $k \geq 2$ and $(i_r,i_{r+1}) \in E$ for $1 \leq r \leq k-1$, we call a sequence of edges $((i_1,i_2),\ldots,(i_{k-1},i_k))$ a path from i_1 to i_k . A path from a node to itself is called a directed cycle. Additionally, a directed graph is called a directed acyclic graph (DAG) if it contains no directed cycles. Given a DAG $m = (V,E)$, a node i_2 is called a parent of i_1 if there is an edge from i_2 to i_1. We write $\mathrm{Pa}(i)$ to denote the set of parents of a node i. A node i_2 is called a descendant of i_1 if there is a path from i_1 to i_2, and i_2 is called a nondescendant of i_1 if i_2 is not a descendant of i_1.

Suppose $m = (V,E)$ is a DAG, and $X = \{\xi_1,\ldots,\xi_N\}$ is a random vector that follows a joint probability distribution P. For each i, let ξ_i correspond to the i^{th} node of V. For $A \subset V$, let ξ_A denote the collection of variables $\{\xi_i : i \in A\}$. (In the literature, sometimes this collection is written simply as A. We will occasionally following this convention, but in mathematical expressions about probabilities we usually prefer to distinguish clearly between the set of variables A and their values ξ_A.) In particular, $\xi_{\mathrm{Pa}(i)}$ describes the states of the parents of node i. We say that (m,P) satisfies the Markov condition if each component of X is conditionally independent of the set of all its nondescendants given the set of all its parents. Finally, if (m,P) satisfies the Markov condition, then we say that (m,P) is a Bayesian network, and that m forms a Bayesian network with P. See Neapolitan (2004) for more details.

The independence constraints encoded in a Bayesian network allow for a simplification of the joint probability distribution P which is captured by the factorization theorem (Neapolitan, 2004, Theorem 1.4):

Theorem 1 *If (m,P) satisfies the Markov condition, then P is equal to the product of its conditional distributions of all nodes given the values of their parents, whenever these conditional distributions exist:*

$$P(\xi_1,\ldots,\xi_N) = \prod_{i=1}^{N} P\left(\xi_i|\xi_{Pa(i)}\right).$$

Consider the following example (also see Neapolitan, 2004, Example 2.9). Rewrite the variables $(\xi_1,\xi_2,\xi_3,\xi_4) = (U,Y,Z,W)$. Suppose we have a Bayesian network (m,P) where m is shown in Figure 1 and the distribution P satisfies the conditions presented in Table 1 for some

$P(u_1) = a$	$P(y_1\|u_1) = 1-(b+c)$	$P(z_1\|y_1) = e$	$P(w_1\|z_1) = g$
$P(u_2) = 1-a$	$P(y_2\|u_1) = c$	$P(z_2\|y_1) = 1-e$	$P(w_2\|z_1) = 1-g$
	$P(y_3\|u_1) = b$	$P(z_1\|y_2) = e$	$P(w_1\|z_2) = h$
	$P(y_1\|u_2) = 1-(b+d)$	$P(z_2\|y_2) = 1-e$	$P(w_2\|z_2) = 1-h$
	$P(y_2\|u_2) = d$	$P(z_1\|y_3) = f$	
	$P(y_3\|u_2) = b$	$P(z_2\|y_3) = 1-f$	

Table 1: Constraints on distribution P

Figure 1: The DAG m for our first example.

$0 \leq a,b,c,\ldots,g,h \leq 1$. Note that, due to Theorem 1, the equations in Table 1 fully determine P as a function of $a,b,c,\ldots,g,h$. Further, since (m,P) satisfies the Markov condition, each node is conditionally independent of the set of all its nondescendants given its parents. For example, we see that Z and U are conditionally independent given Y (written $Z \perp\!\!\!\perp U\,|\,Y$). Do these conditional independences entail any other conditional independences, that is, are there any other conditional independences which P must satisfy other than the one based on a node's parents? The answer is positive. For example, if (m,P) satisfies the Markov condition, then

$$P(w|u,y) = \sum_u P(w|z,u,y)P(z|u,y) = \sum_u P(w|z,y)P(z|y) = P(w|y)$$

and hence $W \perp\!\!\!\perp U\,|\,Y$. Explicitly, the notion of "entailed conditional independence" is given in the following definition:

Definition 2 *Let $m = (V,E)$ be a DAG where V is a set of random variables, and let $A,B,C \subset V$. We say that, based on Markov condition, m entails conditional independence $A \perp\!\!\!\perp B\,|\,C$ if $A \perp\!\!\!\perp B\,|\,C$ holds for every $P \in P_m$, where P_m is the set of all probability distributions P such that (m,P) satisfies the Markov condition.*

We say that there is a *direct dependency* between variables A and B in P if A and B are not conditionally independent given any subset of V. Based on the Markov condition, the absence of an edge between A and B implies that there is no direct dependency between A and B. However, the Markov condition is not sufficient to guarantee that the presence of an edge means direct dependency. In general, given a Bayesian network (m,P), we would want an edge in m to mean there is a direct dependency. In this case the DAG would become what it is naturally expected to be—a graphical representation of the structure of relationships between variables. The faithfulness condition as defined below indeed reflects this.

Definition 3 *We say that a Bayesian network (m,P) satisfies the faithfulness condition if, based on the Markov condition, m entails all and only the conditional independences in P. When (m,P)*

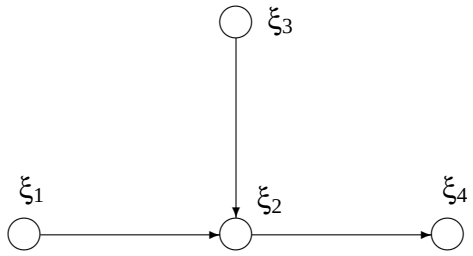

Figure 2: A second example of a DAG.

satisfies the faithfulness condition, we say that m and P are faithful to each other and we say m is a perfect map of P.

It is easy to see that the Bayesian network (m,P), where m is shown in Figure 1 and P satisfies the constraints in Table 1, does not satisfy the faithfulness condition. Indeed, Table 1 implies that $U \perp\!\!\!\perp Z$, but this independence is not entailed by m based on the Markov condition. As shown in Example 2.10 of Neapolitan (2004), the distribution P of this example has no perfect map. However, it is not hard to see that the DAG of Figure 1 is "optimal" in the sense that no DAG with fewer edges forms a Bayesian network with P.

In this paper we concentrate on Bayesian networks over a set of variables $X = \{\xi_1,\ldots,\xi_N\} \sim P$ where each variable takes values from the set $\{1,2\}$. Let m be a DAG with nodes $1,\ldots,N$. The probability distributions in P_m can be parameterized according to the conditional distributions of Theorem 1 as follows. For each node i, let $|\mathrm{Pa}(i)|$ be the number of parents of i and let $q_i(m) = 2^{|\mathrm{Pa}(i)|}$ be the number of possible states of the set of variables $\xi_{\mathrm{Pa}(i)}$. Consider a fixed list of the $q_i(m)$ possible states of $\xi_{\mathrm{Pa}(i)}$. For $j \in \{1,\ldots,q_i(m)\}$, we shall write "$\xi_{\mathrm{Pa}(i)} = j$" to mean that the parents of node i are in the states given by the j^{th} item in the list. For $k = 1,2$ and $j = 1,\ldots,q_i(m)$, we write $\theta_{ijk} = P(\xi_i = k \,|\, \xi_{\mathrm{Pa}(i)} = j)$. Observe that $\theta_{ij2} = 1 - \theta_{ij1}$. We shall write Θ_m to denote the vector of all θ_{ij1}'s for m:

$$\Theta_m \;=\; \big(\theta_{ij1} \,:\, i = 1,\ldots,N,\; j = 1,\ldots,q_i(m)\big) \;\in\; [0,1]^{k_m},$$

where $k_m = \sum_{i=1}^{N} q_i(m)$. Then each Θ_m in $[0,1]^{k_m}$ determines a probability measure $P = P_{\Theta_m}$ such that (m,P) is Bayesian network; and conversely, if (m,P) is a Bayesian network, then $P = P_{\Theta_m}$ for some $\Theta_m \in [0,1]^{k_m}$.

To illustrate this notation, consider the DAG m in Figure 2. Here, $\mathrm{Pa}(1) = \emptyset = \mathrm{Pa}(3)$, $\mathrm{Pa}(2) = \{1,3\}$, and $\mathrm{Pa}(4) = \{3\}$, and so $q_1(m) = 2^0 = q_3(m)$, $q_2(m) = 2^2$, and $q_4(m) = 2^1$. We could fix the list of possible states of $\xi_{\mathrm{Pa}(4)}$ to be "1,2", and the list for $\xi_{\mathrm{Pa}(2)}$ to be "(1,1), (1,2), (2,1), (2,2)" (with the understanding that the ordering is (ξ_1,ξ_3)). For the latter list, we have for example

$$\theta_{231} \;=\; P\big(\xi_2 = 1 \,|\, \xi_{\mathrm{Pa}(2)} = 3\big) \;=\; P\big(\xi_2 = 1 \,|\, (\xi_1,\xi_3) = (2,1)\big)\,.$$

Since $\mathrm{Pa}(3) = \emptyset$, $P(\xi_{\mathrm{Pa}(3)} = 1) = 1$, and θ_{311} is simply $P(\xi_3 = 1)$. We can write

$$\Theta_m \;=\; (\theta_{111},\theta_{211},\theta_{221},\theta_{231},\theta_{241},\theta_{311},\theta_{411},\theta_{421}),$$

and $k_m = 1+4+1+2 = 8$.

Let $D = D_n = \{X_1, ..., X_n\}$ be fully observed data of size n generated according to Pr, and let N_{ijk} be the number of cases in the database D such that node i takes value k while its parent set $\xi_{\mathrm{Pa}(i)}$ takes the values corresponding to j.

A *probabilistic model* $\mathcal{M}$ for a random vector $X = (\xi_1, \ldots, \xi_N)$ is a set of possible joint probability distributions of its components. If the probability distribution P is a member of a model $\mathcal{M}$, we say P is *included* in $\mathcal{M}$. Let m be a DAG (V,E). A *Bayesian network model* is a pair (m,F) where F is a set of possible parameter vectors Θ_m: each Θ_m in F determines conditional probability distributions for m, such that the joint probability distribution P_{Θ_m} of X (given by the product of these conditional distributions) satisfies the Markov condition with m. (E.g., for the DAG m of Figure 2, the most general choice of F is $[0,1]^8$, but F could also be a subset of $[0,1]^8$.) For simplicity, we shall usually omit F when referring to a Bayesian network model (m,F). In a given class of models, if $\mathcal{M}_2$ includes the probability distribution P, and if there exists no $\mathcal{M}_1$ (in the class) such that $\mathcal{M}_1$ includes P and $\mathcal{M}_1$ has smaller dimension than $\mathcal{M}_2$, then $\mathcal{M}_2$ is called a *parameter optimal map* of P. (E.g. the DAG of Figure 1 is a parameter optimal map of the distribution P of Table 1.) For the Bayesian network models we shall work with in this paper, the dimension of a model m is $k_m = \sum_{i=1}^{N} q_i(m)$. A detailed discussion of probabilistic model selection in the case of Bayesian networks could be found in Neapolitan (2004).

In order to proceed further we would also need a formal definition of consistency. In this definition we assume that the dimensions of the probabilistic models are well-defined. For a more detailed discussion of the definition of consistency see, for example, Neapolitan (2004), Grünwald (2007) and Lahiri (2001).

Definition 4 *Let D_n be a set of values (data) of a set of n mutually independent random vectors $X_1, \ldots, X_n$, each with probability distribution P. Furthermore, let* `score` *be a scoring criterion over some class of models for the random variables that constitute each vector. We say* `score` *is* ***consistent*** *for the class of models if the following two properties hold:*
1. If $\mathcal{M}_1$ includes P and $\mathcal{M}_2$ does not, then

$$\lim_{n\to\infty} Pr(\texttt{score}(D_n, \mathcal{M}_1) > \texttt{score}(D_n, \mathcal{M}_2)) = 1.$$

2. If $\mathcal{M}_1$ and $\mathcal{M}_2$ both include P and $\mathcal{M}_1$ has smaller dimension than $\mathcal{M}_2$, then

$$\lim_{n\to\infty} Pr(\texttt{score}(D_n, \mathcal{M}_1) > \texttt{score}(D_n, \mathcal{M}_2)) = 1.$$

Additionally, we say that the scoring criterion is ***strongly consistent*** *if, in both cases 1 and 2, it selects the appropriate model almost surely:*

$$Pr(\exists N : \forall n \geq N \;\; \texttt{score}(D_n, \mathcal{M}_1) > \texttt{score}(D_n, \mathcal{M}_2)) = 1.$$

As an example, let m_1 be the DAG of Figure 2, let m_2 be the DAG obtained from m_1 by adding an arc from node 3 to node 4, and let m_0 be the DAG obtained from m_1 by removing the arc from node 2 to node 4. For $i = 0,1,2$, let $\mathcal{M}_i$ be the probabilistic model consisting of all probability distributions with which m_i forms a Bayesian network. Let P be a probability distribution in $\mathcal{M}_1$ such the components of Θ_{m_1} are eight distinct numbers in $(0,1)$. Then $\mathcal{M}_0$ does not contain P (since ξ_4 is not independent of $\{\xi_1, \xi_2, \xi_3\}$), while $\mathcal{M}_1$ and $\mathcal{M}_2$ both contain P, and $\mathcal{M}_1$ has smaller

dimension that $\mathcal{M}_2$. If `score` is consistent, then in a situation with lots of data, `score` will be very likely to rank $\mathcal{M}_1$ over either $\mathcal{M}_0$ or $\mathcal{M}_2$. However, consider an infinite stream of data $X_1, X_2, \ldots$ sampled independently from P. Suppose that after each new observation, we ask `score` to choose among $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$. Consistency says that the expected proportion of `score`'s correct choices tends to 1 as n tends to infinity. But strong consistency says more: if `score` is strongly consistent, then with probability one it will make the correct choice for all but finitely many values of n.

3. Results

In this paper we consider the case of binomial Bayesian networks with independent $Beta(\alpha_{ij1}, \alpha_{ij2})$ priors for the parameters θ_{ij1} (note that $\theta_{ij2} = 1 - \theta_{ij1}$), where $\alpha_{ij1}, \alpha_{ij2} > 0$. We choose the beta family as it is the conjugate prior for the Binomial distribution. According to (1), the value of the Bayesian scoring criterion can be calculated as follows:

$$\begin{aligned} p(D_n|m) &= \int_{\Omega_m} p(D_n|m,\Theta_m)\, p(\Theta_m|m)\, d\Theta_m \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i(m)} \int_0^1 \theta_{ij1}^{N_{ij1}+\alpha_{ij1}-1} (1-\theta_{ij1})^{N_{ij2}+\alpha_{ij2}-1} \frac{1}{Beta(\alpha_{ij1}, \alpha_{ij2})} d\theta_{ij1} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i(m)} \frac{Beta(N_{ij1}+\alpha_{ij1}, N_{ij2}+\alpha_{ij2})}{Beta(\alpha_{ij1}, \alpha_{ij2})} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i(m)} \frac{\Gamma(N_{ij1}+\alpha_{ij1})\Gamma(N_{ij2}+\alpha_{ij2})}{\Gamma(N_{ij1}+N_{ij2}+\alpha_{ij1}+\alpha_{ij2})} \cdot \frac{\Gamma(\alpha_{ij1}+\alpha_{ij2})}{\Gamma(\alpha_{ij1})\Gamma(\alpha_{ij2})}, \end{aligned} \tag{2}$$

which coincides with the well-known formula by Cooper and Herskovits (1992).

Throughout this paper we produce several asymptotic expansions "in probability" and "almost surely", always with respect to our probability measure Pr on Ω. We derive several properties of the marginal likelihood (2). We shall show that, for any model m,

$$\log p(D_n|m) = nC_m + \mathrm{O}(\sqrt{n \log\log n}) \qquad \text{a.s.}, \tag{3}$$

where C_m is a constant independent of n. We strengthen this result by showing how to obtain a positive constant σ_m such that

$$\limsup_{n\to\infty} \frac{\log p(D_n|m) - nC_m}{\sqrt{2n \log\log n}} = \sigma_m \quad \text{and} \quad \liminf_{n\to\infty} \frac{\log p(D_n|m) - nC_m}{\sqrt{2n \log\log n}} = -\sigma_m \quad \text{a.s.} \tag{4}$$

We note that "in probability" versions of the above statements also follow from our methods (as in the proofs of Corollaries 12 and 10):

$$\log p(D_n|m) \;=\; nC_m + O_p(\sqrt{n}), \tag{5}$$

$$\frac{\log p(D_n|m) - nC_m}{\sqrt{n}} \;\xrightarrow{\mathcal{D}}\; N(0, \sigma_m).$$

Additionally, we will be using the approximation of the Bayesian scoring criterion via maximum log-likelihood:

$$\log p(D_n|m) = \log\left(\prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \hat{\theta}_{ij1}^{N_{ij1}}(1-\hat{\theta}_{ij1})^{N_{ij2}}\right) - \frac{1}{2}k_m \log n + O_p(1), \tag{6}$$

where $\hat{\theta}_{ij1}$ is the MLE of θ_{ij1} and $k_m = \sum_{i=1}^{N} q_i(m)$ is the dimension of the model m. This is the efficient approximation of Bayesian score commonly known as BIC which was first derived in Schwarz (1978) for the case of linear exponential families. In Haughton (1988) his result was made more specific and extended to the case of curved exponential families—the type of model that includes Bayesian networks, as is shown in Geiger et al. (2001).

In this work, we attempt to get insight into the rate of convergence of the Bayes factor comparing two models m_1 and m_2. Our result strengthens the well known property of consistency of the Bayesian scoring criterion (e.g., see Chickering, 2002) and is expressed as the following theorem.

Theorem 5 *In the case of a binomial Bayesian network class, for the Bayesian scoring criterion based on independent beta priors, the following two properties hold:*

1. If m_2 includes P and m_1 does not, then there exists a positive constant $C(P, m_1, m_2)$ such that

$$\log \frac{\texttt{score}_B(D_n|m_2)}{\texttt{score}_B(D_n|m_1)} = C(P, m_1, m_2)n + O(\sqrt{n \log\log n}) \qquad \text{a.s.}$$

and

$$\log \frac{\texttt{score}_B(D_n|m_2)}{\texttt{score}_B(D_n|m_1)} = C(P, m_1, m_2)n + O_p(\sqrt{n}).$$

2. If m_1 and m_2 both include P and $\dim m_1 > \dim m_2$ where $\dim m_k = \sum_{i=1}^{N} q_i(m_k)$, $k = 1, 2$, then

$$\log \frac{\texttt{score}_B(D_n|m_2)}{\texttt{score}_B(D_n|m_1)} = \frac{\dim m_1 - \dim m_2}{2} \log n + O(\log\log n) \qquad \text{a.s.}$$

and

$$\log \frac{\texttt{score}_B(D_n|m_2)}{\texttt{score}_B(D_n|m_1)} = \frac{\dim m_1 - \dim m_2}{2} \log n + O_p(1).$$

In particular, the Bayesian scoring criterion is strongly consistent.

It follows from the consistency property of the Bayesian scoring criterion that if P admits a faithful DAG representation, then the limit of the probability that a consistent scoring criterion chooses a model faithful to P, as the size of data approaches infinity, equals 1. Our result in Theorem 5 strengthens this claim as follows:

Corollary 6 *If (m_1, P) satisfies the faithfulness condition and (m_2, P) does not, then with probability 1, $\frac{\texttt{score}_B(D_n|m_1)}{\texttt{score}_B(D_n|m_2)}$ approaches infinity at exponential rate in n when m_2 does not include P, and approaches infinity at polynomial rate in n when m_2 includes P.*

The first result of Theorem 5 is optimal in the following sense:

Theorem 7 *If m_2 includes P and m_1 does not, then there exist $C_{m_1,m_2} > 0$ and $\sigma_{m_1,m_2} > 0$ such that:*

$$\limsup_{n\to\infty} \frac{\log \frac{score_B(D_n|m_2)}{score_B(D_n|m_1)} - nC_{m_1,m_2}}{\sigma_{m_1,m_2}\sqrt{2n\log\log n}} = 1 \qquad a.s.,$$

$$\liminf_{n\to\infty} \frac{\log \frac{score_B(D_n|m_2)}{score_B(D_n|m_1)} - nC_{m_1,m_2}}{\sigma_{m_1,m_2}\sqrt{2n\log\log n}} = -1 \qquad a.s.$$

and also

$$\frac{\log \frac{score_B(D_n|m_2)}{score_B(D_n|m_1)} - nC_{m_1,m_2}}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,\sigma_{m_1,m_2}).$$

The constants C_{m_1,m_2} and σ_{m_1,m_2} from the above theorem could be defined as follows.

Definition 8 *Consider a single observation $X = (\xi_1,\ldots,\xi_N)$ from P. Define*

$$\phi_{ijk}(X) = \begin{cases} \log\theta_{ijk} & \text{if } \xi_i = k \text{ and } \xi_{Pa(i)} = j \\ 0 & \text{otherwise} \end{cases}$$

and define

$$\tau(X,m) = \sum_{i=1}^{N}\sum_{j=1}^{q_i(m)}\sum_{k=1}^{2}\phi_{ijk}(X).$$

Then define C_{m_1,m_2} and σ_{m_1,m_2} respectively to be the mean and the standard deviation of $\tau(X,m_1) - \tau(X,m_2)$. Also define

$$C_{i,m} = \prod_{j=1}^{q_i(m)} \left[\theta_{ij1}^{\theta_{ij1}}(1-\theta_{ij1})^{1-\theta_{ij1}}\right]^{P\left(\xi_{Pa(i)}=j\right)}.$$

Observe that we have

$$\tau(X,m) = \sum_{i=1}^{N}\log\theta_{i,\xi_{\mathrm{Pa}(i)},\xi_i}.$$

We shall show (see Lemma 13) that $\tau(X,m) = \log P(X)$, and (see proof of Lemma 9) that

$$C_{m_1,m_2} = \sum_{i=1}^{N}\log\frac{C_{i,m_2}}{C_{i,m_1}}. \tag{7}$$

Observe that for any i, the constant $C_{i,m}$ depends only on the conditional probabilities $P(\xi_i|\xi_{\mathrm{Pa}(i)})$ of the model m; therefore, if models m_1 and m_2 have the same set of parents of the i^{th} node, then $C_{i,m_1} = C_{i,m_2}$ and the i^{th} term in (7) is zero.

The quantities defined above will be extensively used throughout the Appendix.

4. Conclusion

In this paper we proved the strong consistency property of the Bayesian scoring criterion for the case of binomial Bayesian network models. We obtained asymptotic expansions for the logarithm of Bayesian score as well as the logarithm of the Bayes factor comparing two models. These results are important extensions of the consistency property of the Bayesian scoring criterion, providing insight into the rates at which the Bayes factor favors correct models. The asymptotic properties are found to be independent of the particular choice of beta parameter priors.

The methods we used are different from the mainstream. One typical way to investigate the properties of Bayesian score is to use BIC approximation and hence reduce the problem to investigation of the maximum log-likelihood term. In this paper we use expression (9) where the first term is the log-likelihood evaluated at the true parameter.

If we use the results of Theorem 5 in the approximation of Bayes scoring criterion by BIC (6), we can see that given two models m_1 and m_2, if both of them include the generating distribution P then their maximum log-likelihoods are within $O(\log\log n)$ of each other, and if one of the models does not include P then the maximum log-likelihoods differ by a leading order of $C(P,m_1,m_2)n$. These are the rates obtained by Qian and Field (2002, Theorems 2 and 3) for the case of model selection in logistic regression. This observation advocates for the existence of a unified approach for a very general class of models which can describe the rates at which Bayesian scoring criterion and its approximations favor correct model choices.

Acknowledgments

We would like to thank Hélène Massam, Yuehua Wu, Guoqi Qian, Chris Field, and the referees for their constructive and valuable comments and suggestions. A part of this work was done during the stay of the second author at the Institute of Mathematical Stochastics, University of Göttingen. He is thankful to M. Denker for his hospitality during this visit. This work was supported in part by a Discovery Grant from NSERC Canada (NM), Russian Foundation for Basic Research (DZ)(09-01-91331-NNIO-a, 09-01-00107-a), Russian Science Support Foundation (DZ) and Grant of the President of the Russian Federation (DZ)(NSh-638.2008.1).

Appendix A.

In this section we provide proofs for the basic facts in this paper and for Theorems 5 and 7. Note that part 1 of Theorem 5 follows directly from Theorem 7. In our derivations we first assume that the parameter prior of every node follows a flat $Beta(1,1)$ distribution. At the end, we shall show how the results can be extended to the case of general beta priors.

We will start from the expression for the marginal likelihood (2). Noticing that $Beta(x+1,y+1) = \left[(x+y+1)\binom{x+y}{x}\right]^{-1}$ we obtain the expression for the Bayesian scoring criterion via binomial coefficients:

$$p(D_n|m) = \left[\prod_{i=1}^{N}\prod_{j=1}^{q_i(m)}(N_{ij1}+N_{ij2}+1)\binom{N_{ij1}+N_{ij2}}{N_{ij1}}\right]^{-1}. \tag{8}$$

Let $P(k,n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. Substituting $\binom{n}{k} = \frac{P(k,n,\theta)}{\theta^k(1-\theta)^{n-k}}$ into (8) with θ taken as θ_{ij1} we obtain an expression for $\log p(D_n|m)$, which will be the fundamental core of our proof:

$$\log p(D_n|m) = \sum_{i=1}^{N}\sum_{j=1}^{q_i(m)}\left[\log\left(\theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}\right) - \log P(N_{ij1}, N_{ij1}+N_{ij2}, \theta_{ij1}) - \log(N_{ij1}+N_{ij2}+1)\right]. \quad (9)$$

The rest of the Appendix is organized as follows. In Lemma 9 we derive the law of iterated logarithm for the first term of (9) by using the function $\tau(X,m)$ introduced in Definition 8. Lemma 11 states asymptotic expansions of each of three terms in (9) hence providing us with an opportunity to get an expansion for $p(D_n|m)$. Asymptotic expressions (3), (4) and (5) are immediate consequences of this lemma. Lemma 13 establishes a fundamental result regarding the log-likelihood evaluated at the true parameter value. It is followed by the proofs of Theorems 5 and 7.

Lemma 9 *Recall the notation of Section 2 and Definition 8. For model m, let* $T(m) = T_n(m) = \sum_{i=1}^{N}\sum_{j=1}^{q_i(m)}\log\left(\theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}\right)$. *Then the following laws of the iterated logarithm hold almost surely:*

$$\limsup_{n\to\infty}\frac{T(m)-n\sum_{i=1}^{N}\log C_{i,m}}{\sigma_m\sqrt{2n\log\log n}} = 1, \qquad \liminf_{n\to\infty}\frac{T(m)-n\sum_{i=1}^{N}\log C_{i,m}}{\sigma_m\sqrt{2n\log\log n}} = -1, \quad (10)$$

$$\limsup_{n\to\infty}\frac{[T(m_1)-T(m_2)]-nC_{m_1,m_2}}{\sigma_{m_1,m_2}\sqrt{2n\log\log n}} = 1, \qquad \liminf_{n\to\infty}\frac{[T(m_1)-T(m_2)]-nC_{m_1,m_2}}{\sigma_{m_1,m_2}\sqrt{2n\log\log n}} = -1. \quad (11)$$

Proof It is not difficult to see that $T(m) = \sum_{r=1}^{n}\tau(X_r,m)$ and

$$E(\tau(X,m)) = \sum_{i=1}^{N}\sum_{j=1}^{q_i(m)}\sum_{k=1}^{2}P\left(\xi_{\mathrm{Pa}(i)}=j\right)\theta_{ijk}\log\theta_{ijk} = \sum_{i=1}^{N}\log C_{i,m}.$$

By the law of the iterated logarithm applied to $T(m)$ we conclude that

$$\limsup_{n\to\infty}\frac{T(m)-n\sum_{i=1}^{N}\log C_{i,m}}{\sigma_m\sqrt{2n\log\log n}} = 1,$$

where σ_m is the standard deviation of $\tau(X,m)$. Further, applying the law of the iterated logarithm to $T(m_1)-T(m_2)$, we obtain the equalities (11) where σ_{m_1,m_2} is the standard deviation of $\tau(X,m_1)-\tau(X,m_2)$. ■

Corollary 10 *The following expressions hold:*

$$T(m_1) - T(m_2) \;=\; nC_{m_1,m_2} + O(\sqrt{n\log\log n}) \qquad a.s., \tag{12}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{q_i(m)} \log\left(\theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}\right) \;=\; n\sum_{i=1}^{N}\log C_{i,m} + O_p(\sqrt{n}), \tag{13}$$

$$\frac{[T(m_1) - T(m_2)] - nC_{m_1,m_2}}{\sqrt{n}} \;\xrightarrow{\mathcal{D}}\; N(0,\sigma_{m_1,m_2}). \tag{14}$$

Proof Obviously, (12) is a direct consequence of (11). Applying the central limit theorem to $T(m)$ and $T(m_1) - T(m_2)$ in the proof of Lemma 9 instead of the law of the iterated logarithm we obtain (13) and (14). ■

Lemma 11 *The following asymptotic expansions hold:*

$$\sum_{i=1}^{N}\sum_{j=1}^{q_i(m)} \log\left(\theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}\right) \;=\; n\sum_{i=1}^{N}\log C_{i,m} + O(\sqrt{n\log\log n}) \qquad a.s.,$$

$$N_{ij1} + N_{ij2} + 1 \;=\; n\left[P\left(\xi_{Pa(i)} = j\right) + o(1)\right] \qquad a.s., \tag{15}$$

$$\log P(N_{ij1}, N_{ij1} + N_{ij2}, \theta_{ij1}) \;=\; -\frac{1}{2}\log n + O(\log\log n) \qquad a.s., \tag{16}$$

$$\log P(N_{ij1}, N_{ij1} + N_{ij2}, \theta_{ij1}) \;=\; -\frac{1}{2}\log n + O_p(1). \tag{17}$$

Proof The first expression follows from (10). Further, note, that each of the variables N_{ijk} is a sum of i.i.d. Bernoulli variables. Based on the law of the iterated logarithm for the number of successes in n Bernoulli trials, as $n \to \infty$

$$N_{ijk} \;=\; n\theta_{ijk}P(\xi_{\mathrm{Pa}(i)} = j) + O(\sqrt{n\mathrm{loglog}n}) \qquad \text{a.s.}, \tag{18}$$

which immediately implies (15). Additionally, using the central limit theorem instead of the law of the iterated logarithm we obtain:

$$N_{ijk} \;=\; n\theta_{ijk}P(\xi_{\mathrm{Pa}(i)} = j) + O_p(\sqrt{n}). \tag{19}$$

Next, we will be using the following version of Local De Moivre-Laplace theorem (see for example p. 46 of Chow and Teicher 1978):

If $n \to \infty$ and $k = k_n \to \infty$ are such that $x_k n^{-\frac{1}{6}} \to 0$, where $x_k = \frac{k-np}{\sqrt{np(1-p)}}$, then

$$P(k,n,p) \;=\; \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)} + \mathrm{o}(1)}. \tag{20}$$

According to the law of the iterated logarithm, $x_k = \frac{k-np}{\sqrt{np(1-p)}} = \mathrm{O}(\sqrt{\log\log n})$ a.s., for the case where k is the number of successes in n i.i.d. Bernoulli trials of probability p. Notice that any such x_k satisfies the condition $x_k n^{-\frac{1}{6}} \to 0$. Therefore we can use (20) to approximate the binomial probability in (9), specifically:

$$\begin{aligned}&\log P(N_{ij1}, N_{ij1}+N_{ij2}, \theta_{ij1}) \\ &\quad = -\log\sqrt{2\pi(N_{ij1}+N_{ij2})\theta_{ij1}(1-\theta_{ij1})} - \frac{(N_{ij1}-(N_{ij1}+N_{ij2})\theta_{ij1})^2}{2(N_{ij1}+N_{ij2})\theta_{ij1}(1-\theta_{ij1})} + \mathrm{o}(1) \quad \text{a.s.} \quad (21)\end{aligned}$$

The first term in this expansion can be simplified based on (15):

$$-\log\sqrt{2\pi(N_{ij1}+N_{ij2})\theta_{ij1}(1-\theta_{ij1})} \;=\; -\frac{1}{2}\log n + O(1) \qquad \text{a.s.} \qquad (22)$$

Applying (18) to the second term we conclude that as $n \to \infty$:

$$-\frac{(N_{ij1}-(N_{ij1}+N_{ij2})\theta_{ij1})^2}{2(N_{ij1}+N_{ij2})\theta_{ij1}(1-\theta_{ij1})} \;=\; O(\log\log n) \qquad \text{a.s.} \qquad (23)$$

Now, (21) could be simplified further based on (22) and (23) to obtain (16). Finally, we can prove (17) analogously to (16) by using (19) instead of (18). Therefore the proof of the lemma is complete. ■

Now it is easy to derive the expansions announced in Sect. 3.

Corollary 12 *Properties (3), (4) and (5) of the marginal likelihood $P(D_n|m)$ hold.*

Proof Using (10), (15) and (16) in (9) and denoting $C_m \stackrel{def}{=} \sum_{i=1}^{N} \log C_{i,m}$ we get (4). Further, (3) is a direct consequence of (4). Finally, (5) can be proved by substituting (15), (13) and (16) into (9). ■

Lemma 13 *Suppose the probability distribution P is a member of the model m. Let $T(m) = \sum_{i=1}^{N}\sum_{j=1}^{q_i(m)} \log\left(\theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}\right)$ for model m. Then $T(m) = \log P(D_n)$.*

Proof Since P is a member of the model m, we know that (m,P) satisfies the Markov condition. Therefore, by the factorization theorem (Theorem 1), we obtain

$$P(D_n) \;=\; \prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \theta_{ij1}^{N_{ij1}}(1-\theta_{ij1})^{N_{ij2}}$$

and the result follows. ■

Next, we will prove Theorems 5 and 7. Note that part 1 of Theorem 5 directly follows from Theorem 7.

Geiger et al. (2001) showed that each of the competing Bayesian network models m could be represented as a C^∞ connected manifold of dimension $\sum_{i=1}^N q_i(m)$ embedded in $\mathcal{R}^N$. In order to keep the notation simple we will be denoting this manifold as m. Every probability distribution for a finite sample space belongs to an exponential family. Therefore, there exists a set of ζ_i, $i = 1,2,\ldots,$ i.i.d. observations from an underlying distribution P_{θ_0} belonging to a full exponential family in standard form with densities $f(\zeta,\theta) = \exp(\zeta\theta - b(\theta))$ with respect to a finite measure on $\mathcal{R}^N$, with $\theta \in \Theta$ the natural parameter space, such that

$$\log\left(\prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \hat{\theta}_{ij1}^{N_{ij1}}(1-\hat{\theta}_{ij1})^{N_{ij2}}\right) = n \sup_{\phi\in m\cap\Theta}(Y_n\phi - b(\phi)), \tag{24}$$

where $Y_n = (1/n)\sum_{i=1}^n \zeta_i$.

Theorem 2.3 of Haughton (1988) provides an expansion of the logarithm of the Bayesian scoring criterion via maximum log-likelihood and, together with (24), guarantees (6). It follows from (3), (6) and (24) that

$$n \sup_{\phi\in m\cap\Theta}(Y_n\phi - b(\phi)) = C_m n + O_p(\sqrt{n\log\log n}). \tag{25}$$

Suppose m_2 includes P and m_1 does not. In this case, Haughton (1988, p.346) guarantees that as $n \to \infty$, we have

$$Pr\left(\sup_{\phi\in m_1\cap\Theta}(Y_n\phi - b(\phi)) + \varepsilon < \sup_{\phi\in m_2\cap\Theta}(Y_n\phi - b(\phi))\right) \to 1$$

for some $\varepsilon > 0$, and by (25) we obtain $C_{m_1} < C_{m_2}$. Hence, $\sum_{i=1}^N \log\frac{C_{i,m_2}}{C_{i,m_1}} > 0$. Now, the result of Theorem 7 can be obtained by using (15), (11) and (16) in (9), and by using (15), (14) and (16) in (9).

Now, suppose both m_1 and m_2 include the true distribution P and $k_{m_2} < k_{m_1}$. For part 2 of Theorem 5, direct application of Lemma 13, (16) and (15) provides the "almost surely" result, while Lemma 13, (17) and (15) prove the "in probability" result.

Finally, we shall show that the results of Theorems 5 and 7 hold for the case of general beta priors. It is not difficult to see that Stirling's approximation implies

$$\lim_{z\to\infty} z^{b-a}\frac{\Gamma(z+a)}{\Gamma(z+b)} = 1. \tag{26}$$

Denote as ψ_1 the flat $Beta(1,1)$ system of priors and denote as ψ_2 the system which, for each parameter θ_{ij1}, assumes the distribution $Beta(\alpha_{ij1},\alpha_{ij2})$, where $\alpha_{ij1},\alpha_{ij2} > 0$. It follows from (2) and (26) that:

$$\begin{aligned}
\frac{p(m|D,\psi_2)}{p(m|D,\psi_1)} &= \prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \frac{Beta(N_{ij1}+\alpha_{ij1}, N_{ij2}+\alpha_{ij2})}{Beta(N_{ij1}+1, N_{ij2}+1)} \cdot \frac{1}{Beta(\alpha_{ij1},\alpha_{ij2})} \\
&= \prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \frac{\Gamma(N_{ij1}+\alpha_{ij1})\Gamma(N_{ij2}+\alpha_{ij2})\Gamma(N_{ij1}+N_{ij2}+2)}{\Gamma(N_{ij1}+1)\Gamma(N_{ij2}+1)\Gamma(N_{ij1}+N_{ij2}+\alpha_{ij1}+\alpha_{ij2})} \cdot \frac{1}{Beta(\alpha_{ij1},\alpha_{ij2})} \\
&\sim \prod_{i=1}^{N}\prod_{j=1}^{q_i(m)} \frac{N_{ij1}^{\alpha_{ij1}-1} N_{ij2}^{\alpha_{ij2}-1}}{(N_{ij1}+N_{ij2})^{\alpha_{ij1}+\alpha_{ij2}-2}} \cdot \frac{1}{Beta(\alpha_{ij1},\alpha_{ij2})}.
\end{aligned}$$

Therefore, using (18) we can conclude that there exists a constant $c > 0$ such that:

$$\lim_{n\to\infty} \frac{p(m|D,\psi_2)}{p(m|D,\psi_1)} = c \quad \text{a.s.},$$

which implies that the results of Theorems 5 and 7 extend to the case of general beta parameter priors.

References

David M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Yuan S. Chow and Henry Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, New York, 1978.

Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001.

Peter D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

Parhasarathi Lahiri. *Model selection*. Institute of Mathematical Statistics, Beachwood, Ohio, 2001.

Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

Guoqi Qian and Chris Field. Law of iterated logarithm and consistent model selection criterion in logistic regression. *Statistics and Probability Letters*, 56(1):101, 2002.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.