Evaluating Bayesian Networks by Sampling with Simplified Assumptions

Saaid Baraty*, Dan A. Simovici*

*University of Massachusetts Boston
Computer Science Department,
Boston, Massachusetts 02125
e-mail{sbaraty,dsim}@cs.umb.edu,

Abstract. The most common fitness evaluation for Bayesian networks in the presence of data is the Cooper-Herskovitz criterion. This technique involves massive amounts of data and, therefore, expansive computations. We propose a cheaper alternative evaluation method using simplified assumptions which produces evaluations that are strongly correlated with the Cooper-Herskovitz criterion.

1 Introduction

We investigate the problem of constructing a Bayesian network for a composite phenomenon $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ where $\mathbf{u}_i$ for $1 \leq i \leq n$ are discrete random variables representing the state assignment of the attributes of $\mathbf{U}$. To accomplish this, we start from a data multiset $\mathcal{D} = \{t_1, t_2, \ldots, t_m\}$ where an n-ary tuple t_i is an instance of the event $\mathbf{U}$. We refer to this multiset as *evidence data set* (data set for short).

A number of assumptions are necessary for deriving a measure for evaluating the fitness of a Bayesian network structure (BNS) for a training data set. Stronger hypotheses make the evaluation more manageable. On the other hand, the model obtained under weaker assumptions is better capable to be conforming with the underlying true distribution of the problem.

Let $G = (\mathbf{U}, E)$ be a directed acyclic graph having $\mathbf{U}$ as its set of vertices and E as its set of edges, which captures the direct probabilistic dependencies among these variables. Let Θ be the collection of parameters which quantifies the joint probability distribution of $\mathbf{U}$ as specified by G. We denote the set of possible assignments of a random variable $\mathbf{u}_i$ by $\mathrm{Dom}(\mathbf{u}_i) = \{u_i^1, \ldots, u_i^{r_i}\}$. The notion of domain can be extended to sets of variables $\mathbf{V}$ using Cartesian product. If the *set of parent nodes* of $\mathbf{u}_i$ is $\mathsf{Par}_G(\mathbf{u}_i)$, then $\mathrm{Dom}(\mathsf{Par}_G(\mathbf{u}_i) = \{U_i^1, \ldots, U_i^{q_i}\}$. The set of *non-descendants of* $\boldsymbol{u}_i$, $\mathsf{nd}_G(\mathbf{u}_i)$ is the set of all nodes in $\mathbf{U}$ excluding $\mathbf{u}_i$ and all its descendants. When it is clear from the context we drop the subscript G. The pair $\mathcal{B} = (G, \Theta)$ satisfies the *local Markov condition* if $P_{\mathcal{B}}(\mathbf{u}_i|\mathsf{nd}(\mathbf{u}_i)) = P_{\mathcal{B}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i))$ for $1 \leq i \leq n$, where $P_{\mathcal{B}}$ is the probability distribution on $\mathbf{U}$ specified by $\mathcal{B}$. The model $\mathcal{B}$ is a *Bayesian* network if it satisfies the local Markov condition. By the chain rule we have: $P_{\mathcal{B}}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) = \prod_{i=1}^{n} P_{\mathcal{B}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i))$. Therefore if we let $\theta_{ijk} = P(\mathbf{u}_i = u_i^k|\mathsf{Par}(\mathbf{u}_i) = U_i^j)$ and $\theta_{ij\cdot} = (\theta_{ij1}, \ldots, \theta_{ijr_i})$ for $1 \leq i \leq n$, $1 \leq k \leq r_i$ and $1 \leq j \leq q_i$, then the joint probability distribution on $\mathbf{U}$ is specified by $\Theta = \{\theta_{ij\cdot}|1 \leq i \leq n \text{ and } 1 \leq j \leq q_i\}$.

2 A Posterior-based Score with A Reduced Assumptions Set

Cooper and Herskovitz introduced the probability $P(G|\mathcal{D})$ as a measure of assessing the fitness of G as a probabilistic model of $\mathcal{D}$. Since $P(\mathcal{D})$ is constant across different networks, we can work with $P(G, \mathcal{D})$. Let Ω_G be the space of all probability distributions Θ for the structure G. Then,

$$P(G, \mathcal{D}) = \int_{\Omega_G(\Theta)} P(\mathcal{D}|\Theta, G) f(\Theta|G) P(G) d\Theta. \tag{1}$$

Recall that Θ is a collection of distributions $\theta_{ij\cdot} = (\theta_{ij1}, \ldots, \theta_{ij(r_i-1)}, 1 - \sum_{k=1}^{r_i-1} \theta_{ijk})$ for all i and j. The vectors $\theta_{ij\cdot}$ for any $(i,j) \in [1..n] \times [1..q_i]$ must satisfy $\sum_{k=1}^{r_i-1} \theta_{ijk} \leq 1$ and $\theta_{ijk} \geq 0$ for all k. Also, Θ itself, the collection of these random vector variables, can be treated as a random variable. $P(\mathcal{D}|\Theta, G)$ is the conditional probability function of data given (G, Θ), $f(\Theta|G)$ is the conditional density function of Θ given structure G, and $P(G)$ is the prior probability function of structure G. To evaluate this integral a number of assumptions were introduced by Cooper and Herskovits (1993). The *data independence* assumes tuples of $\mathcal{D}$ are independent given the network structure. The *local and global independence (LGI)* assumption requires that $\theta_{ij\cdot}$ is conditionally independent of $\theta_{i'j'\cdot}$ for all $(i,j) \neq (i',j')$ given the structure. Based on the LGI assumption, $\Omega(\Theta)$, the space of possible collections Θ can be written as

$$\Omega_G(\Theta) = \Big\{ \prod_{i=1}^{n} \prod_{j=1}^{q_i} (\theta_{ij1}, \ldots, \theta_{ij(r_i-1)}) \in \mathbb{R}^{r_i-1} \ | \ \sum_{k=1}^{r_i-1} \theta_{ijk} \leq 1 \text{ and } \theta_{ij1}, \ldots, \theta_{ij(r_i-1)} \geq 0 \Big\}$$

and we have $f(\Theta|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} g(\theta_{ij\cdot}|G)$ due to the LGI assumption. Cooper and Herskovits (1993) replace f by the above product in Equality (1). Also, they assume the distribution $g(\theta_{ij\cdot}|G)$ for each i and j is uniform. We refer to this assumption as *second order uniform probability (SOUP)*. Heckerman et al. (1995) introduce the BDe metric which is a posterior-based measure similar to CH metric. They use the LGI assumption and three other assumptions: the *second order Dirichlet probability (SODP)* (suggested but not used in Cooper and Herskovits (1993)), the *parameter modularity* and the *multinomial sample (MS)* assumption. SODP is generalization of SOUP assumption which states that $P(\theta_{ij\cdot}|G)$ follows a *Dirichlet* distribution for all i and j. The multinomial sample assumption asserts that if we define the ordered set $\mathcal{D}_l = \{t_1, \ldots, t_{l-1}\}$ then,

$$P\Big(t_l[\mathbf{u}_i] = u_i^k \ | \ t_l[\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}] = (u_1^{v_1}, \ldots, u_{i-1}^{v_{i-1}}), \mathcal{D}_l, (G, \Theta)\Big) = \theta_{ijk},$$

where $t[\mathbf{V}]$ denotes the restriction of $\mathbf{V} \subseteq \mathbf{U}$ on tuple $t \in \mathcal{D}$ and we have the state assignment $\mathsf{Par}_G(\mathbf{u}_i) = U_i^j$ consistent with $t_l[\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}] = (u_1^{v_1}, \ldots, u_{i-1}^{v_{i-1}})$ and $\theta_{ijk} \in \Theta$. Later, the SODP assumption was replaced with two other assumptions, *likelihood equivalence* and *structure possibility*, which imply the SODP assumption. Note that every probability function $g(\theta_{ij\cdot}|G)$ follows a *Dirichlet* distribution which requires r_i parameters. Thus, for each BNS G we need to specify $\sum_{i=1}^{n} q_i r_i$ parameters and this makes this approach impractical. To overcome this difficulty Heckerman et al. (1995) encoded the prior knowledge into a single Bayesian network referred as (*a prior network*) $\mathcal{B}_{pr} = (G_{pr}, \Theta_{pr})$. Then, they set the *Dirichlet* parameter corresponding to probability distribution component θ_{ijk} to $\alpha_{ijk} = N' \cdot P_{\mathcal{B}_{pr}}(\mathbf{u}_i = u_i^k, \mathsf{Par}_{G_{pr}}(\mathbf{u}_i) = U_i^j)$, where N' is a user given parameter which they refer as *equivalent sample size*. The choice of a values of N' and the collection Θ_{pr} without observing data is arbitrary. We use sampling which enable us to let data shape the distribution of the posterior probability on vectors $\theta_{ij\cdot}$.

In the evaluation of the prior $P(G)$ Cooper and Herskovits (1993) assumed an uniform prior distribution. This and other assumptions are based on parameters that need to be arbitarily specified. Sampling enables us to use data as a substitute for strong assumptions or domain knowledge in determining the parameters of the second order probability distribution and the prior probability $P(G)$.

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two disjoint samples from $\mathcal{D}$. We evaluate $P(G|\mathcal{S}_1, \mathcal{S}_2)$ as a measure of fitness of BN structure G. Since $P(\mathcal{S}_1, \mathcal{S}_2)$ does not depend on the specific BNS we can drop it and instead compute $P(G, \mathcal{S}_1, \mathcal{S}_2)$. Note that by chain rule $P(G, \mathcal{S}_1, \mathcal{S}_2) = P(\mathcal{S}_1|G, \mathcal{S}_2) \cdot P(G|\mathcal{S}_2) \cdot P(\mathcal{S}_2)$. If we sample consistently across different structures, then $P(\mathcal{S}_2)$ is constant and can be dropped. Therefore, we adopt $P(\mathcal{S}_1|G, \mathcal{S}_2) \cdot P(G|\mathcal{S}_2)$ as a relative measure of fitness of structures for a data set $\mathcal{D}$. If we repeat the process of sampling, we can extend our measure to

$$\left(\prod_{q=1}^{k} P(\mathcal{S}_{2q-1}|G, \mathcal{S}_{2q}) \cdot P(G|\mathcal{S}_{2q})\right)^{\frac{1}{k}},$$

where $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{2k}$ are samples from $\mathcal{D}$ where $\mathcal{S}_{2q-1} \cap \mathcal{S}_{2q} = \emptyset$ for each q. We refer to this measure as *k-sample validation of structure G for data set $\mathcal{D}$* and denote it by $\mathsf{SAMP}_k(G, \mathcal{D})$.

Let $\mathcal{S} = \{t_1, \ldots, t_a\}$ and $\mathcal{S}'$ be two disjoint samples of $\mathcal{D}$. The first term of $\mathsf{SAMP}_k(G, \mathcal{D})$ can be written as

$$P(\mathcal{S}|G, \mathcal{S}') = \int_{\Omega_G(\Theta)} P(\mathcal{S}|\Theta, G, \mathcal{S}') f(\Theta|G, \mathcal{S}') d\Theta. \tag{2}$$

Let $d = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ be a topological order of nodes of G which represents expert prior knowledge of the domain. Denote by $\mathsf{n}_{\mathcal{S}}(t)$ the number of occurrences of tuple t in $\mathcal{S}$ and let $\gamma_{ijk}(\mathcal{S}) = \left|\{t \in \mathcal{S} \mid t[\{\mathbf{u}_i\}] = u_i^k \wedge t[\mathsf{Par}(\mathbf{u}_i)] = U_i^j\}\right|$ and $\gamma_{ij\cdot}(\mathcal{S}) = \sum_{k=1}^{r_i} \gamma_{ijk}(\mathcal{S})$. Since the attributes of $\mathcal{D}$ are discrete, we have

$$P(\mathcal{S}|\mathcal{B}) = \prod_{l=1}^{a} P(t_l|\mathcal{S}^l, \Theta, G) = \prod_{l=1}^{a}\prod_{i=1}^{n} P(\mathbf{u}_i = t_l[\mathbf{u}_i]|\mathbf{U}_i = t_l[\mathbf{U}_i], \mathcal{S}^l, \Theta, G) = \prod_{l=1}^{a}\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{r=1}^{r_i} \theta_{ijr}^{\lambda_{lijr}},$$

where the first equality is by the chain rule and $\mathcal{S}^l = (t_1, \ldots, t_{l-1})$, the second equality is by assuming MS assumption and $\mathbf{U}_i = (\mathbf{u}_1, \ldots, \mathbf{u}_{i-1})$ and $\lambda_{lijr} = 1$ if $t_l[\mathbf{u}_i] = u_i^r \in \mathrm{Dom}(\mathbf{u}_i)$ and $t_l[\mathsf{Par}_G(\mathbf{u}_i)] = U_i^j \in \mathrm{Dom}(\mathsf{Par}_G(\mathbf{u}_i))$ and $\lambda_{lijr} = 0$ otherwise. Since $\sum_{l=1}^{a} \lambda_{lijr} = \gamma_{ijr}(\mathcal{S})$, we have

$$P(\mathcal{S}|\Theta, G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{r=1}^{r_i} \theta_{ijr}^{\gamma_{ijr}(\mathcal{S})}. \tag{3}$$

Then,

$$P(\mathcal{S}|\Theta, G, \mathcal{S}') = \frac{P(\mathcal{S} \cup \mathcal{S}'|\Theta, G)}{P(\mathcal{S}'|\Theta, G)} = \frac{\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{r=1}^{r_i} \theta_{ijr}^{\gamma_{ijr}(\mathcal{S}\cup\mathcal{S}')}}{\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{r=1}^{r_i} \theta_{ijr}^{\gamma_{ijr}(\mathcal{S}')}} = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{r=1}^{r_i} \theta_{ijr}^{\gamma_{ijr}(\mathcal{S})}. \tag{4}$$

For the second term of right hand side of Equality (2) we have

$$f(\Theta|\mathcal{S}', G) = \frac{P(\mathcal{S}'|\Theta, G) f(\Theta|G)}{\int_{\Omega_G(\Theta)} P(\mathcal{S}'|\Theta, G) f(\Theta|G) d\Theta} \tag{5}$$

We assume the SOUP hypothesis and set each $g(\theta_{ij\cdot}|G) = (r_i - 1)!$. The posterior probability of Θ is conditioned on G in presence of sample $\mathcal{S}'$, as shown in Equality (2). This approach is is different from the one used in Cooper and Herskovits (1993) where SOUP hypothesis has been applied directly without intervention of sample data. Then, we have

$$\int_{\Omega_G(\Theta)} P(\mathcal{S}'|\Theta, G) f(\Theta|G) d\Theta = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left((r_i - 1)! \cdot \frac{\prod_{r=1}^{r_i} \gamma_{ijr}(\mathcal{S}')!}{(\gamma_{ij\cdot}(\mathcal{S}') + r_i - 1)!} \right),$$

from Equality (3) and SOUP , LGI , and from a result from Jeffreys and Jeffreys (1988) (see pages 468-470 of this reference). Thus, from the previous equalities and from (3) and (5) we have,

$$f(\Theta|\mathcal{S}', G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \Gamma(\gamma_{ij\cdot}(\mathcal{S}') + r_i) \prod_{r=1}^{r_i} \frac{\theta_{ijr}^{\gamma_{ijr}(\mathcal{S}')}}{\Gamma(\gamma_{ijr}(\mathcal{S}') + 1)}, \tag{6}$$

where Γ is Euler's function. Combining Equalities (2), (4) and (6) we obtain

$$P(\mathcal{S}|G, \mathcal{S}') = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\gamma_{ij\cdot}(\mathcal{S}') + r_i)}{\Gamma(\gamma_{ij\cdot}(\mathcal{S} \cup \mathcal{S}') + r_i)} \cdot \prod_{r=1}^{r_i} \frac{\Gamma(\gamma_{ijr}(\mathcal{S} \cup \mathcal{S}') + 1)}{\Gamma(\gamma_{ijr}(\mathcal{S}') + 1)},$$

To approximate the quantity $P(G|\mathcal{S})$ we use a slight variation of a measure called the *distribution distortion* introduced in Baraty and Simovici (2009). Here we want to evaluate the conditional independency captured by local Markov condition according to data, that is, we want to assess to what degree the conditions $f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{nd}(\mathbf{u}_i)) = f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i))$ holds for $1 \leq i \leq n$, where $f_{\mathcal{S}}$ is the *frequency function relative to* sample $\mathcal{S} \subseteq \mathcal{D}$. To achieve this, we measure the divergence of the set of probability distributions $f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{nd}(\mathbf{u}_i) = U)$ from the set of probability distributions $f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i) = U[\mathsf{Par}(\mathbf{u}_i)])$ for all i and $U \in \mathrm{Dom}(\mathsf{nd}(\mathbf{u}_i))$.

Definition 2.1 The *local Markov divergence of the fork structure at node* $\boldsymbol{u}_i$ *of* G according to sample $\mathcal{S}$, denoted by $\mathsf{LMD}_{\mathcal{S}}^{G}(\mathbf{u}_i)$, is the number

$$\sum_{U} f_{\mathcal{S}}(\mathsf{nd}(\mathbf{u}_i) = U) \cdot \mathsf{KL}\big[f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{nd}(\mathbf{u}_i) = U), f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i) = U[\mathsf{Par}(\mathbf{u}_i)])\big],$$

where the sum extends over all $U \in \mathrm{Dom}(\mathsf{nd}(\mathbf{u}_i))$. Here $\mathsf{KL}[\mathbf{p}, \mathbf{q}]$ is the *Kullbach-Leibler* divergence between the probability distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$. □

Let $\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}})$ be the Shannon entropy of the set $\mathcal{S}$ partitioned according to the values of $\mathbf{u}$, and let $\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}}|\pi^{\mathbf{W}})$ be the conditional Shannon entropy of the set $\mathcal{S}$ partitioned according to the values of $\mathbf{u}$, conditioned by the partition of $\mathcal{S}$ according to the assignment of the set of attributes $\mathbf{W}$ (see Baraty and Simovici (2009)).

Theorem 2.2 *For* $1 \leq i \leq n$ *we have* $\mathit{LMD}_{\mathcal{S}}^{G}(\boldsymbol{u}_i) = \mathcal{H}_{\mathcal{S}}(\pi^{\boldsymbol{u}_i}|\pi^{\mathit{Par}(\boldsymbol{u}_i)}) - \mathcal{H}_{\mathcal{S}}(\pi^{\boldsymbol{u}_i}|\pi^{\mathit{nd}(\boldsymbol{u}_i)})$.

Theorem 2.3 $\mathit{LMD}_{\mathcal{S}}^{G}(\boldsymbol{u}_i) = 0$ *if and only if* $f_{\mathcal{S}}(\boldsymbol{u}_i|\mathit{nd}(\boldsymbol{u}_i)) = f_{\mathcal{S}}(\boldsymbol{u}_i|\mathit{Par}(\boldsymbol{u}_i))$.

Theorem 2.2 implies that $0 \leq \mathsf{LMD}_{\mathcal{S}}^{G}(\mathbf{u}_i) \leq \mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i})$. By Theorem 2.3 the smaller the value of $\mathsf{LMD}_{\mathcal{S}}^{G}(\mathbf{u}_i)$ is, the closer is the fork structure at node $\mathbf{u}_i$ to satisfy the local Markov condition according to $\mathcal{S}$. Therefore, the Markov condition is closer to be satisfied according to sample $\mathcal{S}$. On another hand, the closer $\mathsf{LMD}_{\mathcal{S}}^{G}(\mathbf{u}_i)$ is to $\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i})$ the more divergent the two probability distributions $f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{nd}(\mathbf{u}_i) = U)$ and $f_{\mathcal{S}}(\mathbf{u}_i|\mathsf{Par}(\mathbf{u}_i) = U[\mathsf{Par}(\mathbf{u}_i)])$ are for every $\mathbf{U} \in \mathrm{Dom}(\mathsf{nd}(\mathbf{u}_i))$. When $\mathsf{LMD}^G_{\mathcal{S}}(\mathbf{u}_i) = \mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i})$, we have $\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i}|\pi^{\mathsf{Par}(\mathbf{u}_i)}) = \mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i})$ and $\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i}|\pi^{\mathsf{nd}(\mathbf{u}_i)}) = 0$. This means that the set $\mathsf{Par}(\mathbf{u}_i)$ has no prediction capability at all at the node $\mathbf{u}_i$ and the set $\mathsf{nd}(\mathbf{u}_i)$ has a perfect predication capability on $\mathbf{u}_i$. Let $\mathsf{BNS}(\mathbf{U})$ be the set of all possible Bayesian structures on set of attributes $\mathbf{U}$. Define $P(G|\mathcal{S})$ as

$$P(G|\mathcal{S}) = \frac{\sum_{i=1}^n \left(\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i}) - \mathsf{LMD}^G_{\mathcal{S}}(\mathbf{u}_i)\right)}{\sum_{G' \in \mathsf{BNS}(\mathbf{U})} \sum_{i=1}^n \left(\mathcal{H}_{\mathcal{S}}(\pi^{\mathbf{u}_i}) - \mathsf{LMD}^{G'}_{\mathcal{S}}(\mathbf{u}_i)\right)}.$$

Using the previous evaluations, $\mathsf{SAMP}_k(G, \mathcal{D})$ can be written as

$$\left(\prod_{q=1}^k P(\mathcal{S}_{2q-1}|G, \mathcal{S}_{2q}) \cdot P(G|\mathcal{S}_{2q})\right)^{\frac{1}{k}} = \left(\prod_{q=1}^k \frac{\sum_{s=1}^n \left(\mathcal{H}_{\mathcal{S}_{2q}}(\pi^{\mathbf{u}_s}) - \mathsf{LMD}^G_{\mathcal{S}_{2q}}(\mathbf{u}_s)\right)}{\sum_{G' \in \mathsf{BNS}(\mathbf{U})} \sum_{s=1}^n \left(\mathcal{H}_{\mathcal{S}_{2q}}(\pi^{\mathbf{u}_s}) - \mathsf{LMD}^{G'}_{\mathcal{S}_{2q}}(\mathbf{u}_s)\right)}\right.$$

$$\left.\cdot \prod_{j=1}^{q_i} \frac{\Gamma(\gamma_{ij\cdot}(\mathcal{S}_{2q}) + r_i)}{\Gamma(\gamma_{ij\cdot}(\mathcal{S}_{2q-1} \cup \mathcal{S}_{2q}) + r_i)} \prod_{r=1}^{r_i} \frac{\Gamma(\gamma_{ijr}(\mathcal{S}_{2q-1} \cup \mathcal{S}_{2q}) + 1)}{\Gamma(\gamma_{ijr}(\mathcal{S}_{2q}) + 1)}\right)^{\frac{1}{k}}.$$

If we consistently sample the data across different structures, we can drop the constant entities with respect to BNS G and assuming $P_G^{\mathcal{S}_{2q}} = \sum_{s=1}^n (\mathcal{H}_{\mathcal{S}_{2q}}(\pi^{\mathbf{u}_s}) - \mathsf{LMD}^G_{\mathcal{S}_{2q}}(\mathbf{u}_s))$ we set,

$$\mathsf{SAMP}_k(G, \mathcal{D}) = \left(\prod_{q=1}^k P_G^{\mathcal{S}_{2q}} \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\gamma_{ij\cdot}(\mathcal{S}_{2q}) + r_i)}{\Gamma(\gamma_{ij\cdot}(\mathcal{S}_{2q-1} \cup \mathcal{S}_{2q}) + r_i)} \prod_{r=1}^{r_i} \frac{\Gamma(\gamma_{ijr}(\mathcal{S}_{2q-1} \cup \mathcal{S}_{2q}) + 1)}{\Gamma(\gamma_{ijr}(\mathcal{S}_{2q}) + 1)}\right)^{\frac{1}{k}}.$$

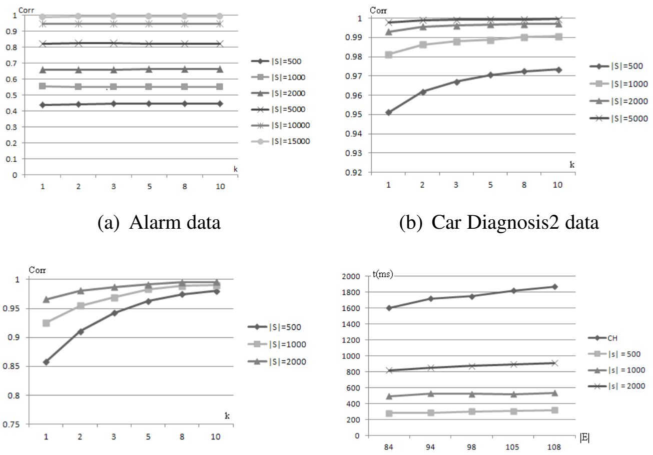
3 Experimental Results and Conclusions

We conducted experiments on three well-known structures G_{AM}, G_{CAR} and G_{NC} for domains Alarm, Car Diagnosis2 and Neapolitan Cancer with 37, 18 and 5 nodes respectively. For the first two structures we randomly generated the corresponding probability tables, Θ_{AM} and Θ_{CAR}. Then, based on probability distributions introduced by (G_{AM}, Θ_{AM}) and (G_{CAR}, Θ_{CAR}) we generated data sets of sizes 80000 and 100000 respectively. For the G_{NC} we used its corresponding data set in the literature with 7565 with no missing values.

For each data set we randomly generated a number of structures of different complexities. The number of the edges for these structures ranged from $1 - 10$, $12 - 108$ and $12 - 330$ for NC, CAR and AM data sets respectively.

Figures 1(a), 1(b) and 1(c) show very strong correlations between the CH score and the SAMP score for various values for k. The derived measure is cheaper to compute, since it works with samples much smaller than the entire data.

We introduced a measure based on posterior probability for measuring the fitness of a Bayesian network structure based on data. The conclusion of this work is that our sampling-based scoring is a viable and much cheaper alternative to the CH score. The fact that we use sampling to reduce the set of assumptions and we get a very strong correlation between two measure confirms that the SOUP and uniform distribution on $P(G)$ are safe assumptions and do not distort the search.



(a) Alarm data

(b) Car Diagnosis2 data

(c) Neapolitan Cancer data

(d) time comparison diagram

FIG. 1 – *Correlations between* $\log(\mathit{SAMP}_k(G, \mathcal{D}))$ *and* $\log(CH)$ *and time in ms needed for computing* $\log(\mathit{SAMP}_1)(G, CAR)$ *and CH scores*

References

Baraty, S. and D. A. Simovici (2009). Edge evaluation in Bayesian network structures. In *Proceedings of the 8th Australian Data Mining Conference*, Melbourne, pp. 193–201.

Cooper, G. F. and E. Herskovits (1993). A Bayesian method for the induction of probabilistic networks from data. Technical Report KSL-91-02, Stanford University, Knowledge System Laboratory.

Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pp. 197–243.

Jeffreys, H. and B. S. Jeffreys (1988). *Dirichlet Integrals*. Cambridge, UK: Cambridge University Press.

Résumé

L'évaluation qualitative la plus connue des réseaux Bayesiens en présence de données est le critère Cooper-Herskovitz. Cette technique implique des quantités massives de données donc, par conséquent, des nombreux calculs. Nous proposons une méthode d'évaluation plus efficace utilisant des suppositions simplifiées et qui produit des évaluations fortement corrélées avec le critère Cooper-Herskovitz.