Bayesian Analysis (2008) **3**, Number 3, pp. 615–624

Exact distribution theory for belief net responses

Peter M. Hooper*

Abstract. Bayesian belief networks provide estimates of conditional probabilities, called query responses. The precision of these estimates is assessed using posterior distributions. This paper discusses two claims and a conjecture by Kleiter (1996) concerning the exact posterior distribution of queries. The two claims provide conditions where a query has an exact beta distribution. The first claim is clarified by the following generalization. Assuming a BDe prior and complete data, a query has the same distribution under equivalent network structures. If the query can be represented as a network parameter under an equivalent structure, it must then have a beta distribution. Kleiter's second claim is contradicted by a counter-example. His conjecture, concerning finite mixtures of beta distributions, is also disproved.

Keywords: Bayesian network, BDe prior, belief network, beta distribution, Dirichlet distribution, query response.

1 Introduction

Bayesian networks (belief networks) are concise models of joint probability distributions. These models can be used to estimate the probability of an event of interest given partial information. As a simple example, consider three binary variables: A indicates the presence a_1 or absence a_2 of a medical condition while B and C each indicate the presence/absence of a symptom. Suppose the two symptoms are known to occur independently of one another among those who have the medical condition and also among those who do not; i.e., B and C are conditionally independent given A. This assumption is represented by the network $\{B \leftarrow A \rightarrow C\}$, where A is called the parent of B and C. Suppose we want to estimate the probability that a subject has the medical condition given that the subject exhibits both symptoms. Such probabilities are called query responses, or simply queries. Put $\theta_a = P\{A = a\}$, $\theta_{b|a} = P\{B = b \,|\, A = a\}$, etc. Applying Bayes Theorem, we have

$$P\{A = a_1 \,|\, B = b_1, C = c_1\} = \frac{\theta_{a_1}\theta_{b_1|a_1}\theta_{c_1|a_1}}{\theta_{a_1}\theta_{b_1|a_1}\theta_{c_1|a_1} + \theta_{a_2}\theta_{b_1|a_2}\theta_{c_1|a_2}}. \tag{A.1}$$

The θ probabilities are the parameters in the model. These parameters are estimated using data from a sample of subjects, then expression (A.1) is applied to estimate the query.

The precision of estimates is usually evaluated via the Bayesian paradigm; i.e., prior

*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada, mailto:hooper@stat.ualberta.ca

 DOI:10.1214/08-BA324

information is combined with data to obtain a posterior distribution for the parameters. The usual prior distributions assume that vectors of θ probabilities (summing to one) are independent Dirichlet random vectors. If we have complete data (i.e., all three variables are observed for each subject in the sample) then the posterior distribution has the same form as the prior. In this case, each θ parameter has a beta distribution. The joint distribution of the parameters determines a distribution for each query. This distribution is usually not analytically tractable, but can often be approximated to an acceptable degree of accuracy by a beta distribution. Kleiter (1996) describes a stochastic simulation technique to carry out this approximation. Van Allen, Singh, Greiner, and Hooper (2008) present a more direct approach, using a delta-rule approximation of the query variance, and prove that the beta approximation is asymptotically valid.

Kleiter (1996) also presents several claims concerning the exact distribution of queries. The present paper provides a critical analysis of this work. One claim is generalized using properties of BDe priors and equivalent network structures. This generalization clarifies how the choice of a prior distribution affects whether a query has an exact beta distribution. A counter-example is constructed for a second claim. The query in expression (A.1) typically does not have a beta distribution, contrary to Kleiter's Theorem 6. A conjecture concerning finite mixtures of beta distributions is also shown to be false. The paper is organized as follows: section 2 introduces notation and assumptions, section 3 presents the generalization, and section 4 discusses Kleiter's work.

2 Notation and assumptions

Let $(\mathcal{V}, \mathcal{A})$ denote a Bayesian network, where $\mathcal{V}$ is a set of random variables (nodes) and $\mathcal{A}$ is a directed acyclic graph (DAG) structure encoding dependence assertions; i.e., a node is conditionally independent of its non-descendants given its parents. Denote variables by upper-case roman letters, realized values by lower-case, and vectors by boldface. Vectors (i.e., sets of variables listed in some order) are represented by concatenation; e.g., $\boldsymbol{Y} = AB$ and $\boldsymbol{y} = ab$. Vectors are sometimes treated as sets of variables; e.g., $A \in \boldsymbol{Y}$. Domains of network variables and vectors are denoted by $\mathcal{D}$; e.g., $\mathcal{D}_{\boldsymbol{Y}} = \mathcal{D}_A \times \mathcal{D}_B$. All such domains are assumed to be finite. When a vector is expressed in terms of subvectors, say $\boldsymbol{X} = \boldsymbol{ABC}$, it is implicitly assumed that the subvectors are pairwise disjoint.

Let $\boldsymbol{X}$ be a vector consisting of all variables in $\mathcal{V}$. The joint distribution of $\boldsymbol{X}$ is determined as a product of conditional probabilities of variables C given parents $\mathrm{pa}(C) = \boldsymbol{F}$. These probabilities, denoted $\theta_{c|\boldsymbol{f}}$, are often presented in two-way tables (CPtables $\boldsymbol{\theta}_{C|\boldsymbol{F}}$), with rows indexed by $\mathcal{D}_{\boldsymbol{F}}$ and columns indexed by $\mathcal{D}_C$. Denote CPtable rows by $\boldsymbol{\theta}_{C|\boldsymbol{f}}$ and let $\boldsymbol{\Theta} = \langle \boldsymbol{\theta}_{C|\mathrm{pa}(C)}, C \in \mathcal{V} \rangle$ be a vector comprising all CPtable parameters. We express uncertainty about parameters by modeling $\boldsymbol{\Theta}$ as a random vector. Uncertainty propagates to query responses; i.e., the probability of a hypothesis $\boldsymbol{H} = \boldsymbol{h}$ given evidence $\boldsymbol{E} = \boldsymbol{e}$. Queries are denoted

$$q_{\boldsymbol{h}|\boldsymbol{e}} = q_{\boldsymbol{h}|\boldsymbol{e}}(\boldsymbol{\Theta}) = P(\boldsymbol{H} = \boldsymbol{h} \,|\, \boldsymbol{E} = \boldsymbol{e}, \boldsymbol{\Theta}). \tag{A.2}$$

We allow $\boldsymbol{E}$ = the empty set, writing $q_{\boldsymbol{h}} = P(\boldsymbol{H} = \boldsymbol{h} \,|\, \boldsymbol{\Theta})$. The notation in (A.2) emphasizes the fact that a query is a function of $\boldsymbol{\Theta}$, hence a random variable.

The posterior distribution of $q_{\boldsymbol{h}|\boldsymbol{e}}$ depends on whether one assumes that the data supporting the posterior distribution of $\boldsymbol{\Theta}$ include the partial observation $\boldsymbol{E} = \boldsymbol{e}$. The inclusion of $\boldsymbol{E}$ is appropriate when making a prediction based on observed evidence, but not when making an inference about a probability of merely potential interest. Our focus here is on the inferential problem. The addition of a partial observation typically has little effect on the query distribution, and has no effect at all when $q_{\boldsymbol{h}|\boldsymbol{e}}$ can be represented as a CPtable parameter under an equivalent network structure; see Section 3. There is a well-known formula that involves both conditioning frameworks: the mean of the query distribution (with $\boldsymbol{\Theta}$ conditioned on $\boldsymbol{E}$) is obtained by replacing parameters by their expected values (not conditioned on $\boldsymbol{E}$) before evaluating the query; i.e., $\mathcal{E}\{q_{\boldsymbol{h}|\boldsymbol{e}}(\boldsymbol{\Theta}) \,|\, \boldsymbol{E} = \boldsymbol{e}\} = q_{\boldsymbol{h}|\boldsymbol{e}}(\mathcal{E}\{\boldsymbol{\Theta}\})$. This result is implicit in Spiegelhalter and Lauritzen (1990) and explicit in Cooper and Herskovits (1992).

Now consider three assumptions; see Table 1 for an example.

Dirichlet prior. The prior distribution is assumed to have the usual properties: different CPtables are independent (global independence), rows within a CPtable are independent (local independence), and each row has a Dirichlet distribution:

$$\boldsymbol{\theta}_{C|\boldsymbol{f}} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{C|\boldsymbol{f}}) \text{ for each } \boldsymbol{f} \in \mathcal{D}_{\mathrm{pa}(C)}, \tag{A.3}$$

where $\boldsymbol{\alpha}_{C|\boldsymbol{f}}$ is a vector of positive weights $\alpha_{c|\boldsymbol{f}}$. It is convenient to indicate summation by replacing an index with a dot. The sum $\alpha_{\cdot|\boldsymbol{f}} = \sum_c \alpha_{c|\boldsymbol{f}}$ provides a measure of precision concerning the parameters, since (A.3) implies that each parameter has a beta distribution

$$\theta_{c|\boldsymbol{f}} \sim \mathrm{Beta}(\alpha_{c|\boldsymbol{f}}, \alpha_{\cdot|\boldsymbol{f}} - \alpha_{c|\boldsymbol{f}}) \tag{A.4}$$

with mean $\mu = \alpha_{c|\boldsymbol{f}}/\alpha_{\cdot|\boldsymbol{f}}$ and variance $\mu(1-\mu)/(1+\alpha_{\cdot|\boldsymbol{f}})$. If C is binary, then (A.4) is equivalent to (A.3) since the probabilities sum to one. The stronger Dirichlet assumption (A.3) is needed for general results concerning variables with finite domain.

BDe constraints. This assumption places restrictions on the prior weights so that prior information is equivalent to a weighted likelihood from a (possibly fictitious) sample of complete observations. Assume there exists a vector $\boldsymbol{\alpha}_{\boldsymbol{X}} = \langle \alpha_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}} \rangle$ of positive weights such that the prior Dirichlet weights $\alpha_{c|\boldsymbol{f}}$ are obtained by summing $\alpha_{\boldsymbol{x}}$ over all $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$ with C and $\mathrm{pa}(C)$ fixed at c and $\boldsymbol{f}$. We refer to the $\alpha_{\boldsymbol{x}}$ as BDe weights, to distinguish them from the Dirichlet weights $\alpha_{c|\boldsymbol{f}}$. In practice, BDe weights might be specified as $\alpha_{\boldsymbol{x}} = \sum w_i I(\tilde{\boldsymbol{y}}_i = \boldsymbol{x})$, where $w_i > 0$, I is the 0/1 indicator function, and each $\tilde{\boldsymbol{y}}_i \in \mathcal{D}_{\boldsymbol{X}}$ is a (fictitious) complete observation. The BDe weights need not be uniquely determined from the Dirichlet weights unless $\boldsymbol{X} = \boldsymbol{A}B$ with $\mathrm{pa}(B) = \boldsymbol{A}$. BDe constraints were originally formulated to facilitate the learning of network structure (Heckerman, Geiger and Chickering, 1995).

Complete data. The posterior distribution is obtained from a sample of n complete observations, where $n = 0$ is allowed. Let $n_{c\boldsymbol{f}}$ denote the number of cases with $C = c$ and $\mathrm{pa}(C) = \boldsymbol{f}$. Given a Dirichlet prior and complete data, the posterior distribution

Table 1: Example of prior and posterior weights, both BDe and corresponding Dirichlet. The DAG structure is $\mathcal{A} = \{B \leftarrow A \rightarrow C\}$ with A and C binary, B ternary. The prior assigns non-uniform weights for B, indicating an expectation that b_1 is most likely and b_3 is least likely, with total weight equivalent to a sample of six observations. The posterior weights are consistent with a sample of 100 observations. Sample frequencies are given by the difference in BDe weights (posterior minus prior). If the arc $B \rightarrow C$ were to be added, then $\alpha_{c|a}$ would be replaced by $\alpha_{c|ab} = \alpha_{abc}$.

Prior	α_{abc}				$\alpha_a = \alpha_{a\cdot\cdot}$		$\alpha_{b\|a} = \alpha_{ab\cdot}$				$\alpha_{c\|a} = \alpha_{a\cdot c}$		
	a_1c_1	a_1c_2	a_2c_1	a_2c_2	a_1	a_2		b_1	b_2	b_3		c_1	c_2
b_1	.75	.75	.75	.75	3.0	3.0	a_1	1.5	1.0	0.5	a_1	1.5	1.5
b_2	.50	.50	.50	.50			a_2	1.5	1.0	0.5	a_2	1.5	1.5
b_3	.25	.25	.25	.25									
Posterior	α_{abc}				$\alpha_a = \alpha_{a\cdot\cdot}$		$\alpha_{b\|a} = \alpha_{ab\cdot}$				$\alpha_{c\|a} = \alpha_{a\cdot c}$		
	a_1c_1	a_1c_2	a_2c_1	a_2c_2	a_1	a_2		b_1	b_2	b_3		c_1	c_2
b_1	4.75	1.75	35.75	15.75	33.0	73.0	a_1	6.5	21.0	5.5	a_1	13.5	19.5
b_2	7.50	13.50	5.50	5.50			a_2	51.5	11.0	10.5	a_2	47.5	25.5
b_3	1.25	4.25	6.25	4.25									

has the same form as the prior but with weights $\alpha_{c|\boldsymbol{f}}$ replaced by $\alpha_{c|\boldsymbol{f}} + n_{c\boldsymbol{f}}$. If the prior Dirichlet weights satisfy BDe constraints, then the posterior weights do as well. Let $\boldsymbol{y}_i \in \mathcal{D}_{\boldsymbol{X}}$ denote the sample observations. Posterior BDe weights are obtained by replacing the prior BDe weights $\alpha_{\boldsymbol{x}}$ with $\alpha_{\boldsymbol{x}} + \sum I(\boldsymbol{y}_i = \boldsymbol{x})$. From now on $\alpha_{\boldsymbol{x}}$ and $\alpha_{c|\boldsymbol{f}}$ denote the posterior BDe and Dirichlet weights, since only the posterior distribution is relevant to our discussion.

3 Query distributions under equivalent DAG structures

The induced dependence model for a DAG structure $\mathcal{A}$ is defined as the subset of triplets $\boldsymbol{ABC} \subseteq \boldsymbol{X}$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ conditionally independent given $\boldsymbol{C}$. Two DAG structures are said to be equivalent if they induce the same dependence model. In general, a query can have different distributions under equivalent structures. However, if the prior distribution under each structure satisfies BDe constraints with the same BDe weight vector $\boldsymbol{\alpha}_{\boldsymbol{X}}$ for all structures, then a query must have the same distribution under equivalent structures. A query $q_{h|\boldsymbol{e}}$ must then have a beta distribution if it can be expressed as a parameter $\theta_{h|\boldsymbol{e}}$ for an equivalent structure. This result is implicit in theory showing that a metric associated with BDe constraints assigns equal support to equivalent DAG structures (Heckerman et al., 1995) and is also closely related to work by Geiger and Heckerman (1997) discussed below. For clarity, a separate proof is provided here. Two preliminary results are needed.

The proof employs a well-known connection between gamma and Dirichlet distribu-

tions; e.g., see Johnson and Kotz (1972), page 271. This connection is also used later in Section 4. A random variable γ has a gamma distribution with shape parameter α and scale parameter β if its density is $\exp(-g/\beta)g^{\alpha-1}\beta^{-\alpha}/\Gamma(\alpha)$ for $g > 0$. The mean and variance of γ are $\alpha\beta$ and $\alpha\beta^2$. If $\beta = 1$, then we write $\gamma \sim \text{Gam}(\alpha)$.

Lemma 1. If $\gamma_1, \ldots, \gamma_n$ are independent with $\gamma_i \sim \text{Gam}(\alpha_i)$, then $\gamma_\cdot := \sum \gamma_i \sim \text{Gam}(\alpha_\cdot)$, $\boldsymbol{\eta} := \langle \gamma_1/\gamma_\cdot, \ldots, \gamma_n/\gamma_\cdot \rangle \sim \text{Dir}(\alpha_1, \ldots, \alpha_n)$, and $\gamma_\cdot$ and $\boldsymbol{\eta}$ are independent. Conversely, if $\boldsymbol{\eta}$ and γ are independent with $\boldsymbol{\eta} := \langle \eta_1, \ldots, \eta_n \rangle \sim \text{Dir}(\alpha_1, \ldots, \alpha_n)$ and $\gamma \sim \text{Gam}(\alpha_\cdot)$, then $\gamma_i := \gamma\eta_i \sim \text{Gam}(\alpha_i)$ with $\gamma_1, \ldots, \gamma_n$ independent.

The proof also uses the following characterization of equivalent DAG structures (Chickering (1995), Heckerman et al. (1995)).

Lemma 2. Given DAG $\mathcal{A}$, put $\boldsymbol{A} = \text{pa}(B)$ and suppose B is a parent of C. The arc $B \to C$ is said to be covered in $\mathcal{A}$ if $\text{pa}(C) = \boldsymbol{A}B$. If $\mathcal{B}$ is obtained from $\mathcal{A}$ by replacing the covered arc $B \to C$ with $B \leftarrow C$, then the two DAGs are equivalent. In general, $\mathcal{A}$ and $\mathcal{B}$ are equivalent if and only if there exists a sequence of DAGs $\mathcal{A}_1, \ldots, \mathcal{A}_m$ over $\mathcal{V}$ such that $\mathcal{A} = \mathcal{A}_1$, $\mathcal{B} = \mathcal{A}_m$, and $\mathcal{A}_{i+1}$ is obtained from $\mathcal{A}_i$ by reversing a single arc that is covered in $\mathcal{A}_i$.

Theorem 1. If we have complete data and Dirichlet priors satisfying BDe constraints with common BDe weight vector $\boldsymbol{\alpha}_{\boldsymbol{X}}$, then a query $q_{\boldsymbol{h}|\boldsymbol{e}}$ has the same distribution under equivalent DAG structures.

Proof. Let $\mathcal{A}$ and $\mathcal{B}$ be equivalent DAG structures. By Lemma 2, it suffices to consider the case where $\mathcal{B}$ is obtained from $\mathcal{A}$ by reversing an arc $B \to C$ that is covered by $\boldsymbol{A}$. Fixing $\boldsymbol{a} \in \mathcal{D}_{\boldsymbol{A}}$, it then suffices to show that $q_{BC|\boldsymbol{a}} := \langle q_{bc|\boldsymbol{a}}, bc \in \mathcal{D}_{BC} \rangle$ has the same joint distribution under both $\mathcal{A}$ and $\mathcal{B}$. Write $\boldsymbol{X} = \boldsymbol{A}BC\boldsymbol{Z}$, where $\boldsymbol{Z}$ includes all other variables. Applying Lemma 1, we represent relevant parameters in terms of independent random variables $\gamma_{bc} \sim \text{Gam}(\alpha_{\boldsymbol{a}bc\cdot})$; i.e., $\theta_{b|\boldsymbol{a}} := \gamma_{b\cdot}/\gamma_{\cdot\cdot}$, $\theta_{c|\boldsymbol{a}b} := \gamma_{bc}/\gamma_{b\cdot}$, $\theta_{c|\boldsymbol{a}} := \gamma_{\cdot c}/\gamma_{\cdot\cdot}$, and $\theta_{b|\boldsymbol{a}c} := \gamma_{bc}/\gamma_{\cdot c}$. This representation satisfies all required joint distributions. Under $\mathcal{A}$, we have $q_{bc|\boldsymbol{a}} = \theta_{b|\boldsymbol{a}}\theta_{c|\boldsymbol{a}b}$ with independent $\boldsymbol{\theta}_{B|\boldsymbol{a}} \sim \text{Dir}(\boldsymbol{\alpha}_{\boldsymbol{a}B\cdot})$ and $\boldsymbol{\theta}_{C|\boldsymbol{a}b} \sim \text{Dir}(\boldsymbol{\alpha}_{\boldsymbol{a}bC\cdot})$. Under $\mathcal{B}$, we have $q_{bc|\boldsymbol{a}} = \theta_{c|\boldsymbol{a}}\theta_{b|\boldsymbol{a}c}$ with independent $\boldsymbol{\theta}_{C|\boldsymbol{a}} \sim \text{Dir}(\boldsymbol{\alpha}_{\boldsymbol{a}\cdot C\cdot})$ and $\boldsymbol{\theta}_{B|\boldsymbol{a}c} \sim \text{Dir}(\boldsymbol{\alpha}_{\boldsymbol{a}Bc\cdot})$. Now $q_{bc|\boldsymbol{a}} = \gamma_{bc}/\gamma_{\cdot\cdot}$ and $q_{BC|\boldsymbol{a}} \sim \text{Dir}(\alpha_{\boldsymbol{a}BC\cdot})$ under both $\mathcal{A}$ and $\mathcal{B}$.

Example 1. Consider three equivalent DAG structures obtained via Lemma 2.

$$A \to B \to C \qquad A \leftarrow B \to C \qquad A \leftarrow B \leftarrow C$$

Each structure induces the dependence model with A and C conditionally independent given B. Given common BDe weights α_{abc}, there are three equivalent representations of the joint distribution: $q_{abc} = \theta_a\theta_{b|a}\theta_{c|b} = \theta_{a|b}\theta_b\theta_{c|b} = \theta_{a|b}\theta_{b|c}\theta_c$. The parameters within each product are independent and each parameter has a beta distribution; e.g., $\theta_{b|c} \sim \text{Beta}(\alpha_{\cdot bc}, \alpha_{\cdot\cdot c} - \alpha_{\cdot bc})$. This example can be generalized in several ways: extend the length of the chain, add common parents to the three variables, and add children to any or all variables. There is no restriction on structure among the added parents or children.

Geiger and Heckerman (1997) obtain a more fundamental result that relates beta-

distributed queries directly to structural hypotheses without requiring the assumption of a Dirichlet prior; i.e., they show that global independence, local independence, and invariance to representation of equivalent structures together imply a Dirichlet prior with BDe constraints. For the simplest nontrivial case, consider three assumptions for a network with $\boldsymbol{X} = AB$.

(i) The $|\mathcal{D}_A| + 1$ random vectors $\{q_A, q_{B|a}, a \in \mathcal{D}_A\}$ are mutually independent; the $|\mathcal{D}_B| + 1$ random vectors $\{q_B, q_{A|b}, b \in \mathcal{D}_B\}$ are mutually independent; and (after removing redundancies from variables summing to one) each set of variables has a strictly positive probability density function.

(ii) q_{AB} has a Dirichlet distribution.

(iii) The network has structure $A \to B$, global independence, local independence, and Dirichlet prior with BDe constraints.

Geiger and Heckerman prove the equivalence of (i) and (ii). The equivalence of (ii) and (iii) is well-known and easily derived; e.g., follow the proof of Theorem 1 above. These equivalences can be generalized by replacing $\{A \to B\}$ with a totally connected set $\boldsymbol{C}$ of variables covered by common parents $\boldsymbol{D}$. The assumption that $q_{\boldsymbol{C}|\boldsymbol{d}}$ has a Dirichlet distribution is equivalent to generalizations of (i) and (iii); see Theorem 3 in Geiger and Heckerman (1997).

4 Discussion of Theorems 5, 6, and 7 in Kleiter (1996)

With BDe constraints, information about all variables is characterized in terms of complete "data" (real and/or fictitious). Kleiter (1996) introduces similar constraints to express this idea for subsets of network variables. Suppose we have a Dirichlet prior and complete data. Let A be a parent of C. Write $\text{pa}(C) = A\boldsymbol{BD}$ and $\text{pa}(A) = \boldsymbol{BE}$, where $\boldsymbol{B}$, $\boldsymbol{D}$, and $\boldsymbol{E}$ are pairwise disjoint and may be empty. Vector $\boldsymbol{B}$ represents parents of both A and C. Vector $\boldsymbol{E}$ cannot include C. We have $\boldsymbol{\theta}_{C|a\boldsymbol{bd}} \sim \text{Dir}(\boldsymbol{\alpha}_{C|a\boldsymbol{bd}})$ and $\boldsymbol{\theta}_{A|\boldsymbol{be}} \sim \text{Dir}(\boldsymbol{\alpha}_{A|\boldsymbol{be}})$. The variable C is said to be a *natural child* of A if

$$\sum\nolimits_{c\boldsymbol{bd}} \alpha_{c|a\boldsymbol{bd}} = \sum\nolimits_{\boldsymbol{be}} \alpha_{a|\boldsymbol{be}} \text{ for all } a \in \mathcal{D}_A. \tag{A.5}$$

The network satisfies natural child constraints if (A.5) holds for all pairs AC with $A \in \text{pa}(C)$.

Theorem 2. BDe constraints imply natural child constraints. The converse holds for networks with just two variables.

Proof. Write $\boldsymbol{X} = A\boldsymbol{BCDEZ}$, where $\boldsymbol{Z}$ includes all other variables. Given BDe constraints with weight vector $\langle \alpha_{a\boldsymbol{bcdez}} \rangle$, we have $\alpha_{c|a\boldsymbol{bd}} = \alpha_{a\boldsymbol{bcd}\cdot\cdot}$ and $\alpha_{a|\boldsymbol{be}} = \alpha_{a\boldsymbol{b}\cdot\cdot\boldsymbol{e}\cdot}$. It follows that both sides of (A.5) equal $\alpha_{a\cdots\cdot}$. If $\boldsymbol{X} = AC$, then "natural child" implies BDe with weights $\alpha_{ac} = \alpha_{c|a}$.

The converse does not hold in general. E.g., suppose $\boldsymbol{X} = ABC$ and $\mathcal{A} = \{A \to B \to C, A \to C\}$. The natural child constraints are $\sum_{ca} \alpha_{c|ab} = \sum_a \alpha_{b|a}$, $\sum_{cb} \alpha_{c|ab} = \alpha_a$, and $\sum_b \alpha_{b|a} = \alpha_a$. Equivalently, setting $\alpha_{abc} := \alpha_{c|ab}$, we have $\alpha_a = \alpha_{a\cdot\cdot}$, $\sum_b \alpha_{b|a} = \alpha_{a\cdot\cdot}$, and $\sum_a \alpha_{b|a} = \alpha_{\cdot b\cdot}$. "Natural child" constrains only the marginal totals of table $\boldsymbol{\alpha}_{B|A}$. BDe constrains the entire table. A similar argument shows that the converse fails for larger networks with complete structure (i.e., where every pair of variables is connected by an arc). An application of Theorem 1 shows that, given complete data, Dirichlet prior, BDe constraints, and complete structure, we have $q_{\boldsymbol{X}} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\boldsymbol{X}})$ and so every query has a beta distribution. This conclusion can fail if BDe is replaced with natural child constraints.

The following example is related to Kleiter's Theorems 5 and 7.

Example 2. Suppose $\boldsymbol{X} = AB$ and $\mathcal{A} = \{A \to B\}$. Assuming complete data and Dirichlet prior (but not necessarily BDe constraints), we have $\boldsymbol{\theta}_A \sim \mathrm{Dir}(\boldsymbol{\alpha}_A)$, $\boldsymbol{\theta}_{B|a} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{B|a})$, all CPtable rows are independent, and $q_{a|b} = \theta_a \theta_{b|a} / \sum_{a'} \theta_{a'} \theta_{b|a'}$. Applying Lemma 1, we represent θ_a as $\gamma_a/\gamma_\cdot$, where the γ_a are independent random variables with $\gamma_a \sim \mathrm{Gam}(\alpha_a)$. We also assume that $\langle \gamma_a \rangle$ is independent of $\boldsymbol{\theta}_{B|A}$. With b fixed, put $\zeta_a = \gamma_a \theta_{b|a}$ so that $q_{a|b} = \zeta_a/\zeta_\cdot$. The ζ_a are independent random variables and the conditional distribution of ζ_a given $\theta_{b|a}$ is gamma with shape parameter α_a and scale parameter $\theta_{b|a}$. The unconditional distribution of ζ_a is thus a continuous scale mixture of gamma distributions. Put $\Delta_a = \alpha_a - \sum_b \alpha_{b|a}$. If $\Delta_a = 0$ (i.e., we have BDe constraints, equivalent to "natural child" here), then the second part of Lemma 1 shows that $\zeta_a \sim \mathrm{Gam}(\alpha_{b|a})$ and consequently $q_{a|b} \sim \mathrm{Beta}(\alpha_{b|a}, \alpha_{b|\cdot} - \alpha_{b|a})$. This is a restatement of Kleiter's Theorem 5. Theorem 1 provides an interpretation of this result: $q_{a|b}$ has a beta distribution because it is a parameter $\theta_{a|b}$ under the equivalent structure $\{A \leftarrow B\}$.

If Δ_a is a positive integer, then the distribution of ζ_a can be expressed as a finite mixture of gamma distributions with fixed scale parameter and varying shape parameter; i.e., $\sum w_k \mathrm{Gam}(\delta_k)$ where the mixing weights w_k are Polya-Eggenberger (PE) probabilities. The shape parameter δ_k varies by increments of 1 from $\alpha_{b|a}$ to $\alpha_{b|a} + \Delta_a$. A derivation of this result is provided in the Appendix. If Δ_a is a nonnegative integer for each $a \in \mathcal{D}_A$, then it follows that the distribution of $q_{a|b}$ is a mixture of beta distributions where the mixing weights are products of PE probabilities.

This beta mixture result constitutes part of Kleiter's Theorem 7. His full Theorem 7 makes a more general claim that the distribution of $q_{a|b}$ is a finite mixture of beta distributions even if some or all Δ_a are negative integers. For negative Δ_a it is claimed that the shape parameter δ_k varies by increments of 1 from $\max\{0, \alpha_{b|a} + \Delta_a\}$ to $\min\{\alpha_{b|a}, \alpha_a\}$. No proofs are given in Kleiter (1996) but further details are provided by Kleiter and Kardinal (1995). Their result for $\Delta_a < 0$ is based on an an additional assumption (an urn model sampling scheme), which is not needed for $\Delta_a > 0$. They conjecture that the result for $\Delta_a < 0$ remains valid without this additional assumption.

Theorem 3 below implies that the beta mixture claim for $\Delta_a < 0$ is in fact not possible given a Dirichlet posterior distribution. This suggests that the urn model sam-

pling scheme may be incompatible with the Dirichlet assumption. Given the connection between Dirichlet and gamma distributions, a finite mixture of beta distributions would imply corresponding finite shape mixtures of gamma distributions, one for each $a \in \mathcal{D}_A$, but this is possible only if the Δ_a are nonnegative integers. The proof of Theorem 3 is given in the Appendix.

Theorem 3. Suppose θ and γ are independent random variables with $\theta \sim \text{Beta}(\alpha_1, \alpha_2)$ and $\gamma \sim \text{Gam}(\alpha_3)$. Put $\zeta = \gamma\theta$ and $\Delta = \alpha_3 - \alpha_1 - \alpha_2$. The distribution of ζ is a finite shape mixture of gamma distributions if and only if Δ is a nonnegative integer. If $\Delta = 0$, then $\zeta \sim \text{Gam}(\alpha_1)$.

Kleiter's Theorem 6 makes the following claim. If $\boldsymbol{B}$ denotes all children of A, each child of A is a natural child with respect to all of its parents, and all CPtable rows are independent with Dirichlet distributions, then $q_{a|\boldsymbol{b}}$ has a beta distribution. The following example shows that the claim is not valid even with "natural child" replaced by the stronger BDe constraints.

Example 3. Let $\boldsymbol{X} = ABC$ and $\mathcal{A} = \{B \leftarrow A \rightarrow C\}$. Assuming complete data, Dirichlet prior, and BDe constraints with weights α_{abc}, we have independent $\boldsymbol{\theta}_A \sim \text{Dir}(\boldsymbol{\alpha}_{A\cdot\cdot})$, $\boldsymbol{\theta}_{B|a} \sim \text{Dir}(\boldsymbol{\alpha}_{aB\cdot})$, and $\boldsymbol{\theta}_{C|a} \sim \text{Dir}(\boldsymbol{\alpha}_{a\cdot C})$. Fixing $bc \in \mathcal{D}_{BC}$ and following the steps in Example 2, we obtain

$$q_{a|bc} \;=\; \theta_a\theta_{b|a}\theta_{c|a} / \sum\nolimits_{a'} \theta_{a'}\theta_{b|a'}\theta_{c|a'} \;=\; \zeta_a\theta_{c|a} / \sum\nolimits_{a'} \zeta_{a'}\theta_{c|a'} \;.$$

The ζ_a are independent random variables, $\zeta_a \sim \text{Gam}(\alpha_{ab\cdot})$, and $\langle\zeta_a\rangle$ is independent of $\langle\theta_{c|a}\rangle$. Put $\Delta_a = \alpha_{ab\cdot} - \alpha_{a\cdot\cdot}$. By Theorem 3, we need $\Delta_a = 0$ in order for $\zeta_a\theta_{c|a}$ to have a gamma distribution. Here we have $\Delta_a < 0$ for all $a \in \mathcal{D}_A$, since all α_{abc} are positive. The distribution of $q_{a|bc}$ is neither beta nor a finite mixture of betas. It is possible for $q_{a|bc}$ to have a beta distribution given certain non-BDe constraints on the weights; e.g., $\alpha_a = \sum_b \alpha_{b|a}$ and $\alpha_{b|a} = \sum_c \alpha_{c|a}$ for all $ab \in \mathcal{D}_{AB}$. Such constraints are unlikely to occur in posterior distributions.

5 Appendix: Proof of Theorem 3 on finite mixtures

The conditional distribution of ζ given θ is gamma with shape parameter α_3 and scale parameter θ, so the marginal density of ζ is

$$f(\zeta) = \int_0^1 \left\{ \frac{1}{\Gamma(\alpha_3)\theta^{\alpha_3}} \zeta^{\alpha_3 - 1} \exp(-\zeta/\theta) \right\} \left\{ \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} \right\} d\theta \;.$$

Make the change of variable $\eta = (1-\theta)/\theta$, so $\theta = (\eta+1)^{-1}$ and $d\theta = -(\eta+1)^{-2}d\eta$. Some algebraic manipulation yields $f(\zeta) = g(\zeta)h(\zeta, \alpha_2, \Delta)$, where $\Delta = \alpha_3 - \alpha_1 - \alpha_2$,

$$\begin{aligned} g(\zeta) &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \zeta^{\alpha_3 - 1} \exp(-\zeta) \;, \\ h(\zeta, \alpha_2, \Delta) &= \int_0^\infty \exp(-\zeta\eta)\, \eta^{\alpha_2 - 1}(\eta+1)^{\Delta} d\eta \;. \end{aligned} \tag{A.6}$$

Now suppose f can be represented as a finite mixture of $\text{Gam}(\alpha_k^*)$ densities with mixing weights w_k. Define moments $M_n(\alpha_3) = E(\gamma^n)$. A comparison of $E(\zeta^n) = E(\theta^n)M_n(\alpha_3)$ with $w_k M_n(\alpha_k^*)$ as $n \to \infty$ shows that each α_k^* must be strictly less than α_3. It follows that, without loss of generality, we may adopt the following parameterization:

$$\begin{aligned} \alpha_k^* &= \alpha_1 + \Delta - \beta_k \,, \\ w_k &= c_k \frac{\Gamma(\alpha_1 + \Delta - \beta_k)}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + \beta_k)}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + \Delta)} \,, \end{aligned} \tag{A.7}$$

with k ranging over a finite set $\mathcal{K}$, $-\alpha_2 < \beta_k < \alpha_1 + \Delta$, $c_k > 0$, and $\sum w_k = 1$. Using this parameterization, the mixture assumption is equivalent to the following representation of h in (A.6):

$$\int_0^\infty \exp(-\zeta\eta)\, \eta^{\alpha_2 - 1}(\eta + 1)^\Delta d\eta = \sum c_k \Gamma(\alpha_2 + \beta_k)\zeta^{-(\alpha_2 + \beta_k)}. \tag{A.8}$$

Now observe that (A.8) is valid if and only if

$$(\eta + 1)^\Delta = \sum c_k \eta^{\beta_k} \quad \text{for all } \eta > 0. \tag{A.9}$$

Integration shows that (A.9) implies (A.8) and the converse follows from uniqueness of the Laplace transform. If Δ is a nonnegative integer, then (A.9) holds with $\mathcal{K} = \{0, \ldots, \Delta\}$, $\beta_k = k$, and binomial coefficients $c_k = \Delta!/\{k!(\Delta - k)!\}$. The weights (A.7) are then Polya-Eggenberger probabilities (Johnson and Kotz, 1977). Otherwise (A.9) cannot hold (see below) and so f cannot be a finite gamma mixture.

If $\Delta < 0$, then (A.9) would imply $c_k = 0$ for $\beta_k \neq 0$. To see this, first note that as $\eta \to 0$ we have $(\eta + 1)^\Delta \to 1$, $\eta^\beta \to 0$ for $\beta > 0$, and $\eta^\beta \to \infty$ at different rates for different $\beta < 0$. Thus c_k must be zero if $\beta_k < 0$. Second, as $\eta \to \infty$ we have $(\eta + 1)^\Delta \to 0$, $\eta^\beta \to 0$ for $\beta < 0$, and $\eta^\beta \to \infty$ at different rates for different $\beta > 0$. Thus c_k must be zero if $\beta_k > 0$.

Suppose Δ is positive and not an integer. If $\beta < 0$ or if β is not an integer then, as $\eta \to 0$, η^β or one of its derivatives becomes unbounded (at different rates for different β) while the function $(\eta + 1)^\Delta$ and each of its derivatives remains bounded. Thus (A.9) would require that all β_k be nonnegative integers. On the other hand, an examination of $(\eta+1)^\Delta \eta^{-\beta}$ as $\eta \to \infty$ shows that the maximum β_k must equal Δ. This contradiction shows that (A.9) cannot hold.

References

Chickering, M. D. (1995). "A transformational characterization of equivalent Bayesian networks." In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (eds. P. Besnard and S. Hanks) San Mateo: Morgan Kaufman, 87–98.

Cooper, G. and Herskovits, E. (1992). "A Bayesian method for the induction of probabilistic networks from data." *Machine Learning* 9: 309–347.

Geiger, D. and Heckerman, D. (1997). "A characterization of the Dirichlet distribution through global and local parameter independence." *The Annals of Statistics* 25: 1344–1369.

Heckerman, D. , Geiger, D., and Chickering, D. M. (1995). "Learning Bayesian networks: the combination of knowledge and statistical data." *Machine Learning* 20: 197–243.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory*. New York: Wiley.

Kleiter, G. D. (1996). "Propagating imprecise probabilities in Bayesian networks." *Artificial Intelligence* 88: 143–161.

Kleiter, G. D. and Kardinal, M. (1995). "A Bayesian approach to imprecision in belief nets." In *Symposia Gaussiana, Proceedings of the 2nd Gauss Symposium, Conference B: Statistical Sciences* (eds. V. Mammitzsch and H. Schneeweiß) Berlin: De Gruyter, 91–105.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990). "Sequential updating of conditional probabilities on directed graphical structures." *Networks* 20: 579–605.

Van Allen, T., Singh, A., Greiner, R. and Hooper, P. M. (2008). "Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference." *Artificial Intelligence* 172: 483–513.

Acknowledgments

The author wishes to thank Russ Greiner, Byron Schmuland, and two referees for helpful comments. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.