On Parameter Priors for Discrete DAG Models

Dmitry Rusakov and **Dan Geiger**
{rusakov,dang}@cs.technion.ac.il
Computer Science Department
Technion, Israel.

Abstract

We investigate parameter priors for discrete DAG models. It was shown in previous works that a Dirichlet prior on the parameters of a discrete DAG model is inevitable assuming global and local parameter independence for all possible complete DAG structures. A similar result for Gaussian DAG models hinted that the assumption of local independence may be redundant.

Herein, we prove that the local independence assumption is necessary in order to dictate a Dirichlet prior on the parameters of a discrete DAG model. We explicate the minimal set of assumptions needed to dictate a Dirichlet prior, and we derive the functional form of prior distributions that arise under the global independence assumption alone.

1 Introduction

A directed graphical model is a representation of a family of joint probability distributions for a collection of random variables via a Directed Acyclic Graph. Each node in the DAG corresponds to a random variable, and the lack of an edge between two nodes represents a conditional independence assumption. A specific joint probability distribution can be represented by a given directed graphical model by specifying the values for the set of associated parameters. The DAG along with such a distribution is called a *Bayesian network*. Graphical models and Bayesian networks have been extensively studied in AI, Statistics, Machine Learning, and in many application areas [2-7,9,11].

Bayesian networks encode a probability distribution with a manageable number of parameters (due to the factorization introduced by underlying graph), thus reducing the complexity of the representation and reducing the complexity of decision making based on this distribution. Bayesian networks are also useful when constructed directly from expert knowledge because they introduce cause-effect relationships that are intuitive to human experts. These features made Bayesian networks a premier tool for representing probabilistic knowledge and reasoning under uncertainty.

In this paper we focus on *learning*—the process of updating both the parameters and the structure of a Bayesian network based on data. To compute goodness-of-fit of data to a network structure in a closed form, researchers have made a number of assumptions. Among them, global and local parameter independence for all network structures, Dirichlet distribution on network parameters, and some other assumptions [2]. It was later shown that the assumption of global and local parameter independence for all nodes in every complete network structure dictates that the only possible prior parameter distribution for discrete DAG models is a Dirichlet prior [5, 7].

In contrast, in a subsequent work, it was shown that for Gaussian DAG models, which consist of a recursive set of linear regression models, global independence alone dictates that the only feasible parameter prior is the Normal-Wishart distribution, assuming models with at least three nodes [6]. It was thus natural to hypothesize that the proofs for discrete and continuous case can be unified and, as a result, the assumption of local independence will turn out to be redundant also in the characterization of the Dirichlet distribution.

This work shows that, while global parameter independence dictates a Normal-Wishart prior for Gaussian DAG models with more than 3 nodes, global parameters independence alone does not dictate a Dirichlet prior for discrete DAG models with more than 3 nodes. We provide a minimal set of assumptions needed to dictate a Dirichlet prior and, in addition, we specify the class of discrete probability distributions, which is larger than the Dirichlet family, that arise under global independence assumption alone via a solution of a new set of functional equations.

2 DAG Models

A Directed Acyclic Graphical model $m \triangleq m(s, \mathcal{F}_s)$ for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ each associated with a set of possible values D_i, is a set of joint probability distributions with sample space $\mathbf{D} = D_1 \times \ldots \times D_n$ specified via two components: a structure s and a set of local distribution families $\mathcal{F}_s$.

The structure s for $\mathbf{X}$ is a DAG having for every variable X_i in $\mathbf{X}$ a node labeled X_i. We denote the parents of X_i by $\mathbf{Pa}_i^s$. The structure s represents the set of conditional independence assertions, and only these conditional independence assertions, which are implied by a factorization of a joint distribution for $\mathbf{X}$ given by $p(\mathbf{x}) = \prod_{i=1}^n p(x_i|\mathbf{pa}_i^s)$, where $\mathbf{x}$ is a value for $\mathbf{X}$ (an n-tuple), x_i is a value for X_i and $\mathbf{pa}_i^s$ is a value for $\mathbf{Pa}_i^s$. When x_i has no incoming arcs in s (no parents), $p(x_i|\mathbf{pa}_i^s)$ stands for $p(x_i)$. A DAG model is *complete* if it has no missing arcs. Note that any two complete DAG models for $\mathbf{X}$ encode the same set of conditional independence assertion, namely none.

The local distributions are the n conditional and marginal distributions that constitute the factorization of $p(\mathbf{x})$. Each such distribution belongs to the specified family of allowable probability distributions $\mathcal{F}_s$, which depends on a finite set of numerical parameters $\theta_m \in \Theta_m \subseteq \mathbb{R}^k$. The parameters θ_m^i for a local distribution is a set of real numbers that completely determine the functional form of $p(x_i|\mathbf{pa}_i^s)$.

We restrict our discussion to discrete DAG models, where local distributions $p(x_i|\mathbf{pa}_i^s)$ are specified by multinomial parameters $\theta_m^i = \{\theta_{x_i|\mathbf{pa}_i^s} | x_i \in D_i, \mathbf{pa}_i^s \in \mathbf{D}_{\mathbf{Pa}_i^s}\}$, where $\mathbf{D}_{\mathbf{Pa}_i^s}$ is the set of possible values of $\mathbf{Pa}_i^s$. Let θ_m denote $\langle \theta_m^1, \theta_m^2, \ldots, \theta_m^n \rangle$ and let $\Theta_{\mathbf{X}}$ denote the set of *joint multinomial parameters* for $\mathbf{X}$, i.e. $\Theta_{\mathbf{X}} = \{\theta_{\vec{x}} | \vec{x} \in \mathbf{D}\}$.

According to the Bayesian framework, we suppose there exists a prior distribution $p(\Theta_{\mathbf{X}})$. This prior induces the distributions of network parameters for each complete model $p(\theta_m|m)$ via a change of parameters formula, because two complete models with multinomial parameters represent the same set of distributions. We assume the regularity of parameter distributions.

Assumption 1 (Regularity) *The probability distribution functions on joint parameters and corresponding p.d.f.'s on network parameters for all complete models are everywhere positive and twice differentiable.*

This paper investigates the functional from of the prior distributions $p(\Theta_{\mathbf{X}})$ that satisfy the properties of *global* and/or *local* parameter independence. Global parameter independence for one network was introduced in [11] to allow a decomposable prior-to-posterior analysis and global parameter independence for all the networks was introduced in [2] in order to search among candidate models.

Definition *Parameters θ_m of a DAG model m are said to be* globally independent *if $\{\theta_m^i\}_{i=1}^n$ are mutually independent, i.e. $p(\theta_m|m) = \prod_{i=1}^n p(\theta_m^i|m)$.*

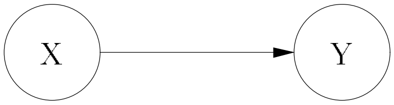
Definition *Parameters θ_m^i of a node X_i of a DAG model $m(s, \mathcal{F}_s)$ are said to be* locally independent *if the subsets $\theta_{X_i|\mathbf{pa}_i^s} = \{\theta_{x_i|\mathbf{pa}_i^s} | x_i \in \{d_i^1, \ldots, d_i^{|D_i|-1}\}\}$ of θ_m^i are mutually independent, i.e. $p(\theta_m^i|m) = \prod_{\mathbf{pa}_i^s \in \mathbf{D}_{\mathbf{Pa}_i^s}} p(\theta_{X_i|\mathbf{pa}_i^s}|m)$ for $1 \leq i \leq n$.*

We say that $p(\Theta_{\mathbf{X}})$ satisfies the *global (or local) parameter independence assumption* if the parameters θ_m are globally (or locally) independent under this distribution for *all complete network structures*, in this case we also say that parameters $\Theta_{\mathbf{X}}$ are globally (or locally) independent.

3 Two Node Networks

We commence by deriving a functional form of globally independent distribution for parameter priors of complete two-node network assuming global parameter independence. The results and techniques developed in this section are the basis for the general result.

Consider the following complete two-node network, with variables X,Y having n and k states respectively.



Since this network is complete it is capable of describing any multinomial distribution of two random variables. Any multinomial distribution, described by a set of parameters $\{\theta_{X=i,Y=j}\}$ (in short denoted by $\{\theta_{ij}\}$), can be described by this network by specifying $\theta_{i\cdot} \triangleq \theta_{X=i,\cdot} = \sum_{j=1}^k \theta_{ij}$ and $\theta_{j|i} \triangleq \theta_{Y=j|X=i} = \theta_{ij}/\theta_{X=i,\cdot}$ for $1 \leq i \leq n$ and $1 \leq j \leq k$.

We are interested in finding a functional form of a prior distributions $p(\{\theta_{ij}\})$ that satisfy a global parameter independence assumption for all complete network for $\{X, Y\}$, namely $X \to Y$ (shown above) and $X \leftarrow Y$. Thus according to our assumption, such distributions should satisfy the following two functional equations, which encode global parameter independence:

$$
\begin{array}{l}
p(\{\theta_{ij}\}) = J_1^{-1} f_1(\{\theta_{i\cdot}\}) g_1(\{\theta_{j|i}\}) \\
p(\{\theta_{ij}\}) = J_2^{-1} f_2(\{\theta_{\cdot j}\}) g_2(\{\theta_{i|j}\})
\end{array} \qquad (1)
$$

where J_1,J_2 are appropriate Jacobians and $\theta_{\cdot j}$, $\theta_{i|j}$ are defined similarly to $\theta_{i\cdot}$ and $\theta_{j|i}$.

We formulate the following theorem that extends the result stated in [5] for two-node DAG models with binary variables.

Theorem 1 *Any probability distribution on $\{\theta_{ij}\}$ that satisfies the regularity assumption (1) and global parameter independence assumption (Equation 1), is of the form*

$$\begin{aligned} p(\{\theta_{ij}\}) = C \left[\prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{\alpha_{ij}}\right] \cdot \\ H\left(\left\{\frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,j}\theta_{i,j+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1\right\}\right) \end{aligned} \tag{2}$$

where α_{ij} are arbitrary positive constants, $H()$ is an arbitrary Lebesgue integrable, everywhere positive and twice-differentiable function of $(n-1)(k-1)$ variables and C is a normalization constant.

Theorem 1 implies that for two-node discrete DAG models global parameter independence assumption alone does not guarantee the Dirichlet distribution of priors. In Section 4 we will prove the similar result for all discrete DAG models. Note, that when H is a constant, as happens if local parameter independence is assumed, then $p(\{\theta_i j\})$ is a Dirichlet distribution [5].

The proof of this theorem is based on the direct solution of a system of functional equations 1. The general approach is given in the following subsections.

3.1 The Functional Equation

Consider two sets of variables $\{y_i | 1 \leq i \leq n-1\}$ and $\{z_{ji} | 1 \leq i \leq n, 1 \leq j \leq k-1\}$. The domain of each of these variables is $(0,1)$. These sets correspond to the sets $\{\theta_{i\cdot}\}$ and $\{\theta_{j|i}\}$ of multinomial parameters discussed above. We define

$$\begin{array}{ll} y_n = 1 - \sum_{i=1}^{n-1} y_i & \\ z_{ki} = 1 - \sum_{j=1}^{k-1} z_{ji}, & 1 \leq i \leq n \\ x_j = \sum_{i=1}^{n} z_{ji} y_i, & 1 \leq j \leq k \\ w_{ji} = \frac{z_{ji} y_i}{x_j} & 1 \leq j \leq k,\ 1 \leq i \leq n. \end{array} \tag{3}$$

Note that $x_k = 1 - \sum_{j=1}^{k-1} x_j$ and $w_{jn} = 1 - \sum_{i=1}^{n-1} w_{ji}$ (for $1 \leq j \leq k$). Here, $\{x_j\}$ corresponds to $\{\theta_{\cdot j}\}$ and $\{w_{ji}\}$ corresponds to $\{\theta_{i|j}\}$. Finally, we let

$$\begin{array}{ll} Y = (y_1, \ldots, y_{n-1}), & Z_i = (z_{1,i}, \ldots, z_{k-1,i}), \\ X = (x_1, \ldots, x_{k-1}), & W_j = (w_{j,1}, \ldots, w_{j,n-1}) \\ Z = (Z_1, \ldots, Z_n), & W = (W_1, \ldots, W_k) \end{array} \tag{4}$$

The functional equation we solve (1) can now be expressed as follows

$$F(Y)g(Z) = G(X)f(W) \tag{5}$$

by absorbing Jacobians appearing in Equation 1 inside the functions F, g, G and f that correspond to f_1, g_1, f_2 and g_2 respectively. Note that the free variables in Equation 5 are $y_1, \ldots, y_{n-1}$ and z_{ji}, $1 \leq j \leq k-1$, $1 \leq i \leq n$. All other variables appearing in Equation 5 are defined by Equations 3 and 4.

The solution of Equation 5, which is outlined in the next subsection, is based on the technique of reducing functional equations to partial differential equations ([1], page 324). Similar technique was used in [5].

3.2 Outline of Solution of Equation 5

The solution of Equation 5 relies on the fact that distribution functions are everywhere positive and twice-differentiable. Thus, it is possible to take the logarithm of the original equation and take the first and second derivatives.

We use the following notations: Let $\hat{h}(x)$ denote $\ln h(x)$ for any positive function $h(x)$. Also let

$$\begin{array}{ll} \hat{F}_i(Y) = \frac{\partial \hat{F}(Y)}{\partial y_i} & 1 \leq i \leq n-1 \\ \hat{g}_{ji}(Z) = \frac{\partial \hat{g}(Z)}{\partial z_{ji}} & \begin{array}{l} 1 \leq i \leq n, \\ 1 \leq j \leq k-1 \end{array} \end{array} \tag{6}$$

and similarly for G and f.

Taking the derivatives of the logarithm of Equation 5 wrt (*with respect to*) y_i and wrt z_{ji}, and pushing the derivatives $\hat{f}_{ji}(Z)$ out of the resulting equations we get (for $1 \leq j \leq k-1$):

$$\begin{aligned} \sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \hat{F}_l(Y) = \\ \sum_{l=1}^{n} \left[\left(\frac{z_{jl}}{x_j} - \frac{z_{kl}}{x_k}\right) \sum_{m=1}^{k-1} z_{ml} \hat{g}_{ml}(Z)\right] \\ + \hat{G}_j(X) - \sum_{l=1}^{n} \frac{z_{jl}}{x_j} \hat{g}_{jl}(Z). \end{aligned} \tag{7}$$

Taking a derivative wrt z_{ji} and substituting $z_{ji} \equiv \frac{1}{k}$ (and thus $x_j \equiv \frac{1}{k}$, $w_{ji} \equiv y_i$) we get (for $1 \leq i \leq n-1$):

$$-\sum_{l=1}^{n-1} y_l \hat{F}_l(Y) + \hat{F}_i(Y) = \frac{1}{y_i} C_i + A \tag{8}$$

where C_i, A are some constants. Solving Equation 8 we get

$$F(Y) = C \prod_{i=1}^{n} y_i^{C_i} \tag{9}$$

where C and C_i are some constants. Similarly, $G(X) = B \prod_{j=1}^{k} x_j^{B_j}$. After substituting the solutions for $F(Y)$ and $G(X)$ into Equation 7 and setting $y_i \equiv \frac{1}{n}$ ($x_j \equiv \frac{1}{n} z_{j\cdot}$ where $z_{j\cdot} = \sum_{i=1}^{n} z_{ji}$), the general solution for $\hat{g}_{ji}(Z)$ is a Dirichlet solution plus the general solution of the following homogeneous first-order partial differential equation:

$$\sum_{l=1}^{n} \left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right) \sum_{m=1}^{k-1} z_{ml} \hat{g}_{ml}(Z)\right] - \sum_{l=1}^{n} \frac{z_{jl}}{z_{j\cdot}} \hat{g}_{jl}(Z) = 0. \tag{10}$$

The general solution of Equation 10 can be shown to be

$$g(Z) = h\left(\left\{\frac{z_{ji}z_{j+1,i+1}}{z_{j+1,i}z_{j,i+1}} \middle| \begin{array}{c} 1 \leq i \leq n-1, \\ 1 \leq j \leq k-1 \end{array}\right\}\right) \qquad (11)$$

where h is an arbitrary everywhere positive, Lebesgue integrable and twice differentiable function. Combining the results of Equation 9 and Equation 11 we conclude the proof of Theorem 1. ■

4 Multiple Node Networks: Globally Independent Parameters

Consider a complete DAG model on n discrete variables: $\mathbf{X} = X_1, \ldots, X_n$, each having $|D_1|, \ldots, |D_n|$ values respectively. In this section we are interested in determining the functional form of distributions on Θ_X that satisfy global parameter independence assumption, i.e. $p(\Theta_\mathbf{X})$ satisfies the following $n!$ functional equations:

$$p(\Theta_\mathbf{X}) = J_I^{-1}\prod_{j=1}^{n} f_{I,j}(\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}), \qquad (12)$$

for all $I = \langle i_1, \ldots, i_n\rangle$ permutations on $\langle 1, \ldots, n\rangle$ where $f_{I,j}()$ are Lebesgue integrable functions that correspond to local parameter distributions. The network parameters $\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}$ are expressed in terms of $\Theta_\mathbf{X}$ and J_I denotes the Jacobian of transformation from the joint parameters to the parameters of the complete Bayesian network with topological order of nodes specified by I. Note, that J_I can be absorbed into $f_{I,j}$, since J_I is a function of $\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}$ (see [7], Theorem 10).

4.1 Useful Lemmas

We present now a set of lemmas that allow the computation of the exact from of globally independent distribution for any set of discrete random variables $\mathbf{X}$.

In order to solve Equation 12 we use Theorem 1. Consider two discrete random variables $\mathbf{Y}_i = \{Y_i, Y\}$, where $Y_i = X_i$ and $Y = X_1 \times \ldots \times X_{i-1} \times X_{i+1} \times \ldots \times X_n$. We claim the following lemma:

Lemma 2 *Given that $p(\Theta_\mathbf{X})$ satisfies the regularity assumption (1), $\Theta_\mathbf{X}$ are globally independent if and only if $\Theta_{\mathbf{Y}_i}$ are globally independent for all $i = 1, \ldots, n$.*

Proof: The 'only if' part of the proof is immediate after noting the correspondence between $\Theta_\mathbf{X}$ and $\Theta_{\mathbf{Y}_i}$. The 'if' part of the proof is done by analyzing the functional form of globally independent distributions for $\Theta_{\mathbf{Y}_i}$, that are obtained using Theorem 1. ■

Now, we apply Theorem 1 for $\mathbf{Y}_i$ and conclude that any $p(\Theta_\mathbf{X})$ that satisfies Equation 12 satisfies the following n equations (for $i = 1, \ldots, n$):

$$p(\Theta_\mathbf{X}) = C_i \prod_{r\in\mathbf{D}} \theta_r^{\alpha_{r,i}} H_i\left(\left\{\frac{\theta_{r_i}\theta_{r_i'''}}{\theta_{r_i'}\theta_{r_i''}}\right\}\right) \qquad (13)$$

where $r_i, r_i', r_i'', r_i''' \in \mathbf{D}$ are subsequent indexes with respect to X_i and $X \setminus X_i$ (analogous to the arguments in Equation 2). I.e., when restricted to X_i: $[r_i]_{-X_i} = [r_i'']_{-X_i} \triangleq a$, $[r_i']_{-X_i} = [r_i''']_{-X_i} \triangleq b$ and $b = a+1$, and when restricted to $X \setminus X_i$: $[r_i]_{-X\setminus X_i} = [r_i']_{-X\setminus X_i} \triangleq c$, $[r_i'']_{-X\setminus X_i} = [r_i''']_{-X\setminus X_i} \triangleq d$ and $d = c+1$. Here $[r]_{-X}$ denote the vector of values of r for nodes $X \subseteq \mathbf{X}$.

Lemma 2 specifies that the set of solutions of Equation 12 is equivalent to the solutions of Equation 13, which are obtained using the following lemma:

Lemma 3 *Consider the following system of m functional equations:*

$$\begin{array}{l} f(x_1, \ldots, x_n) = \sum_{i=1}^n \alpha_{1i}x_i + h_1(\vec{b}_{11}\vec{x}, \ldots, \vec{b}_{1k_1}\vec{x}) \\ f(x_1, \ldots, x_n) = \sum_{i=1}^n \alpha_{2i}x_i + h_2(\vec{b}_{21}\vec{x}, \ldots, \vec{b}_{2k_2}\vec{x}) \\ \vdots \\ f(x_1, \ldots, x_n) = \sum_{i=1}^n \alpha_{mi}x_i \\ \qquad\qquad +h_m(\vec{b}_{m1}\vec{x}, \vec{b}_{m2}\vec{x}, \ldots, \vec{b}_{mk_m}\vec{x}) \end{array} \qquad (14)$$

where $f, h_1, \ldots, h_m$ are unknown functions, α_{ji} are unknown constants and $\vec{b}_{ji}$ are arbitrary (given) n-dimensional vectors. For applications in this paper, $\vec{b}_{ji} \in \{-1, 0, 1\}^n$ and $k_1 = k_2 = \ldots = k_m$.

The general solution for f in Equation 14 is:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} \alpha_i x_i + h(\vec{b}_1\vec{x}, \ldots, \vec{b}_l\vec{x}) \qquad (15)$$

where h is an arbitrary function, $\{\alpha_i\}$ are arbitrary constants and $\vec{b}_1, \ldots, \vec{b}_l$ is the basis of the linear space $\bigcap_{i=1}^m B_i$, where B_i is a linear space spanned by $\vec{b}_{i1}, \ldots, \vec{b}_{ik_i}$.

Since Equations 13 can be transformed to the form of Equation 14 by taking a logarithm of both sides of each equation and changing the variables to $\ln\theta_r$, Lemma 3 provides a powerful tool for solving Equation 13. The proof of this lemma is quite straightforward by changing the variables inside the h-functions in such way that they include $\vec{b}_1\vec{x}, \ldots, \vec{b}_l\vec{x}$.

Application of Lemmas 2 and 3 provides the functional form of globally independent distribution for any specific set of random variables $\mathbf{X}$. However, the exact functional form of a globally independent distribution for a general $\mathbf{X}$ is too cumbersome, so we present the result for binary-values networks only.

4.2 Binary-Valued Networks

The following theorem gives the exact functional form of globally independent prior distributions for binary valued network. This result extends the result stated in [5] for DAG models with two binary variables and demonstrates that global parameter independence assumption alone is not enough to ensure Dirichlet prior for networks of any size (contrary to the Gaussian DAG models, [6]).

Theorem 4 *Any distribution on $\Theta_{\mathbf{X}}$, where $\mathbf{X} = X_1, \ldots, X_n$ are binary random variables, that satisfies regularity (1) and global parameter independence assumptions is of the form*

$$p(\Theta_{\mathbf{X}}) = C \left[\prod_{\vec{x} \in \{0,1\}^n} \theta_{\vec{x}}^{\alpha_{\vec{x}}} \right] h \left(\frac{\prod_{\vec{x} \in A_0} \theta_{\vec{x}}}{\prod_{\vec{x} \in A_1} \theta_{\vec{x}}} \right) \qquad (16)$$

where h is an arbitrary measurable function, $\{\alpha_{\vec{x}}\}$ are arbitrary positive constants and C is a normalization constant. The set A_0 is the set of all binary vectors of length n with even number of "ones" and the set A_1 is the set of all binary vectors of length n with an odd number of "ones".

The full proof, based on Lemmas 2 and 3, is explicated in the full version of this paper [10].

5 Dirichlet Priors: The Minimal Set of Assumptions

We have shown in the previous sections that global parameter independence alone is not enough to ensure a Dirichlet prior on the network parameters. The natural question is: "What is a minimal set of independence requirements that ensure Dirichlet prior?". In this section we give an answer to this question. We start by providing an additional result that links between global parameter independence in various networks.

We say that the parameters of node X_i, θ^i_m, are *globally independent* if $p(\theta_m|m) = p(\theta^i_m|m)p(\theta_m \setminus \theta^i_m|m)$.

Lemma 5 *Let m_1 be an arbitrary complete n-node network with topological order of nodes $X_{i_1}, \ldots, X_{i_n}$, $\{i_1, \ldots, i_n\} = \{1, \ldots, n\}$ and let m_2 be another complete network, with order $X_{j_1}, \ldots, X_{j_n}$ ($\{j_1, \ldots, j_n\} = \{1, \ldots, n\}$). Then given $i_k = j_k$ and $\{i_1, \ldots, i_{k-1}\} = \{j_1, \ldots, j_{k-1}\}$: $\theta^{i_k}_{m_1}$ are globally independent if and only if $\theta^{j_k}_{m_2}$ globally independent.*

Proof: The proof if straightforward using the correspondence between parameters θ_{m_1} and θ_{m_2}. ■

We can now present the second key result of this paper.

Theorem 6 *Let $\mathbf{X} = X_1, \ldots, X_n$ be random variables over $D_1, \ldots, D_n$. Let $m_1(s_1, \mathcal{F}_{s_1})$ be an arbitrary, complete DAG model for $\mathbf{X}$ with topological order of nodes $X_{i_1}, \ldots, X_{i_n}$, $\{i_1, \ldots, i_n\} = \{1, \ldots, n\}$, and let $m_2(s_2, \mathcal{F}_{s_2})$ be another complete DAG model for $\mathbf{X}$, with order $X_{j_1}, \ldots, X_{j_n}$ ($\{j_1, \ldots, j_n\} = \{1, \ldots, n\}$), s.t. $j_n = i_1$. If the parameters of X_{i_1} in m_1 are globally independent, i.e.*

$$p(\theta_{m_1}|m_1) = p(\theta^{i_1}_{m_1}|m_1)p(\theta_{m_1} \setminus \theta^{i_1}_{m_1}|m_1) \qquad (17)$$

and the parameters of X_{j_n} in m_2 are globally and locally independent, i.e.

$$\begin{aligned} p(\theta_{m_2}|m_2) = \\ p(\theta_{m_2} \setminus \theta^{j_n}_{m_2}|m_2) \textstyle\prod_{\mathbf{pa}^{s_2}_{j_n} \in \mathbf{D}_{\mathbf{Pa}^{s_2}_{j_n}}} p(\theta_{X_{j_n}|\mathbf{pa}^{s_2}_{j_n}}|m_2) \end{aligned} \qquad (18)$$

where $\theta_{X_i|\mathbf{pa}^s_i} = \{\theta_{x_i|\mathbf{pa}^s_i}|x_i \in D_i\}$, and $p(\Theta_{\mathbf{X}})$ satisfies Assumption 1, then $p(\Theta_{\mathbf{X}})$ is Dirichlet and this set of conditions is minimal in the sense that the elimination of any one of these two conditions extends the class of admissible priors beyond a Dirichlet distribution.

The theorem states that among the $n!$ sets of global and local parameter independence assumptions used by previous authors, one actually need only two assumptions: global parameter independence for the network parameters for the first node in some complete network, and global and local parameter independence for the same node in other complete network where this node is the last node.

Proof: The minimality of these two assumptions is straightforward, since eliminating any one of them will allow any distribution of the form given by Equation 17 or Equation 18. Since Lemma 5 holds, we can assume that two DAG models under consideration are models with node orders $X_n, X_1, \ldots, X_{n-1}$ and $X_1, \ldots, X_n$ respectively. By treating nodes $X_1, \ldots, X_{n-1}$ as a one super node for a random variable $Y = X_1 \times X_2 \times \ldots \times X_{n-1}$ the problem reduces to determining prior distributions for two-node network with global parameter independence for all configurations and local parameter independence for one last node in one network.

For a two-node network with n and k node-states, Equations 17 and 18 transform to:

$$\begin{aligned} p(\{\theta_{ij}\}) &= f_1(\{\theta_{i\cdot}\}_{i=1}^{n-1}) g_1(\{\theta_{j|i}\}_{i=1,\ldots,n}^{j=1,\ldots,k-1}) \\ p(\{\theta_{ij}\}) &= f_2(\{\theta_{\cdot j}\}_{j=1}^{k-1}) \textstyle\prod_{j=1}^k h_j(\{\theta_{i|j}\}_{i=1}^{n-1}) \end{aligned} \qquad (19)$$

Any solution p that satisfies Equation 19 satisfies also Equation 1 and thus can be written in form given by Theorem 1 (Equation 2). We have

$$\begin{aligned} C \left[\textstyle\prod_{i=1}^n \prod_{j=1}^k \theta_{ij}^{\alpha_{ij}} \right] H \left(\left\{ \tfrac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,i}\theta_{i,j+1}} \Big|_{1 \le i \le n-1}^{1 \le j \le k-1} \right\} \right) \\ = f_2(\{\theta_{\cdot j}\}_{j=1,\ldots,k-1}) \textstyle\prod_{j=1}^k h_j(\{\theta_{i|j}\}_{i=1,\ldots,n-1}) \end{aligned} \qquad (20)$$

Expressing θ_{ij} in terms of $\theta_{\cdot j}$ and $\theta_{i|j}$ and solving for f_2 we get that f_2 is of Dirichlet form. Absorbing free variables inside H and h_j, denoting $\theta_{i|j}$ by w_{ji}, and taking the logarithm, yields (for any $\theta_{\cdot j}$):

$$\hat{H}\left(\left\{\frac{w_{ji}w_{j+1,i+1}}{w_{j+1,i}w_{j,i+1}}|_{1\leq i\leq n-1}^{1\leq j\leq k-1}\right\}\right) = \sum_{j=1}^{k}\hat{h}_j(\{w_{ji}\}_{i=1}^{n-1}) \tag{21}$$

the solution of which is

$$h_j(w_{j1},\ldots,w_{j,n-1}) = \beta_j\prod_{i=1}^{n} w_{ji}^{\beta_{ji}}, \quad 1\leq j\leq k \tag{22}$$

where $\beta_j, \beta_{j1}, \ldots, \beta_{j,n}$ are constants. Combining the results of solution of Equation 19 for f_2 and Equation 22 we conclude the proof. ■

6 Discussion

This paper shows that local parameter independence is essential in the characterization of a Dirichlet prior via discrete DAG models (Section 5, Theorem 6). In addition, the functional form of prior distributions that arise from global parameter independence assumption alone are investigated (Sections 3 and 4, Theorem 4).

Methods for solving functional equations that are developed in this work allow us to compute prior distributions that arise under global parameter independence assumption for any DAG model (and not only for binary variables). However, the explicit general formula for such priors is not compact due to a large number of variables involved. Instead, we have developed a procedure (based on Lemmas 2 and 3) to specify such distribution (in symbolic form) for any specific DAG model (not described here, see [10]).

All the results presented in this paper were achieved under the assumption of local parameter distributions being twice differentiable and everywhere positive. One may hope to derive the properties of twice differentiability and being everywhere positive for probability density functions of Theorem 6 (Equation 19) using the techniques presented in [8], as done in [5, 6].

Another open question is the question of functional form of the prior distribution that arises from local parameter independence assumption alone. In particular, it is unknown (even for two binary variables) if global parameter independence in second condition in Theorem 6 is essential, or it is enough to assume the local independence alone. The integral functional equation that arises from this reduced set of assumptions is of the form (for two binary variables):

$$\begin{aligned} g_0(z_0)g_1(z_1) = \textstyle\int_0^1 & G\left(z_0y + z_1(1-y)\right) \\ & f\left(\tfrac{z_0y}{z_0y+z_1(1-y)}, \tfrac{(1-z_0)y}{1-z_0y-z_1(1-y)}\right)dy \end{aligned} \tag{23}$$

where g_0, g_1, G and f are unknown functions and z_0, z_1, y are variables from $(0,1)$. The general solution for this equation is unknown and the question “Is there any Lebesgue integrable solution that is not of the Dirichlet form?” is open.

References

[1] J. Aczél. *Lectures on Functional Equations and Their Applications.* Academic Press, 1966.

[2] G. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[3] A. Dawid and S. Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[5] D. Geiger and D. Heckerman. A characterization of the dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25(3):1344–1369, 1997.

[6] D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. To appear in Annals of Statistics, 2001.

[7] D. Heckerman, D. Geiger, and D. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[8] A. Járai. Regularity property of the functional equation of the dirichlet distribution. *Aequationes Mathematicae*, 56:37–46, 1998.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[10] D. Rusakov and D. Geiger. On parameter priors for discrete dag models. Technical Report CIS-2000-08, Technion, 2000.

[11] D. Spiegelhalter and S. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.