Bayesian Networks: Model Selection and Applications

research report

Dmitry Rusakov
under the supervision of Assoc. Prof. Dan Geiger

November 29, 2000

Chapter 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 (DAG). In particular, 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 [9-10,15-19,21,33,38].

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 work we focus on *learning* - the process of updating both the parameters and the structure of a Bayesian network based on data. There are two aspects of the learning process: *model selection* - selecting the best structure to fit the observations and *parameter learning* - selecting the best parameters to the chosen structure.

We take the Bayesian approach to model selection, which chooses model M according to the maximum posteriori probability given the observed data D:

$$P(M|D) \propto P(M,D) = P(M)P(D|M) = P(M)\int P(D|M,\theta)P(\theta|M)d\theta$$

where θ denotes the model parameters.

To compute goodness-of-fit ($P(D|M)$) 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, a Dirichlet prior distribution on network parameters, and some other assumptions [9]. 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 [17, 21]. In Chapter 4, we continue this work and explore a minimal set of assumptions needed to dictate a Dirichlet prior.

When parameter prior ($P(\theta|M)$) is unknown, or has a functional form that does not allow closed form computations, one should use the asymptotic methods for evaluation of $P(D|M)$. One such method, now termed *Bayesian Information Criterion* (BIC), was introduced by Schwarz [37] and was proved to be optimal for choosing the model that comes from linear and curved exponential families [37, 20]. Chapter 3 presents the Bayesian Information Criterion for model selection, and proves its optimality for linear exponential families (after Schwarz, [37]). Such families correspond to the undirected graphical models [30]. We intend to extend the result of Schwarz for the class of *Stratified Exponential Families* (SEF), which include the class of directed graphical models with hidden variables and other classes.

This report starts by showing the previous results: Chapter 2 introduces the concepts of DAG models and exponential families, which form the essential basis for understanding and development further results in the field; and Chapter 3 presents the Bayesian Information Criterion for model selection. Our original results on the functional form of parameter priors for discrete DAG models are described in Chapter 4 and future research goals are summarized in Chapter 5.

Chapter 2

Preliminaries

This chapter introduces the concept of DAG models and exponential families of distributions. This concepts are closely dependent, since for the graphical models, various models correspond to a different subfamilies of an exponential family of distributions ([30, 18]).

2.1 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 $\theta_m \in \Theta_m \subseteq \mathbb{R}^k$. (a parametric family). The parameters θ_m^i for a local distribution is a set of real numbers that completely determine the functional form of $p(x_i|\mathbf{pa}_i^s)$.

In discrete DAG models, which describe the distributions on discrete variables $\{X_1, \ldots, X_n\}$, 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 $\mathbf{Pa}_i^s$. In these models, θ_m denotes the set of multinomial parameters for each node $\langle \theta_m^1, \theta_m^2, \ldots, \theta_m^n \rangle$ and $\Theta_{\mathbf{X}}$ denotes the set of *joint multinomial parameters* for $\mathbf{X}$, i.e. $\Theta_{\mathbf{X}} = \{\theta_{\vec{x}} | \vec{x} \in \mathbf{D}\}$.

In Gaussian DAG models, the parameters θ_m^i correspond to the parameters of the linear regression model [16]

$$p(x_i|\mathbf{pa}_i^s, \theta_m^i) = N(x_i|\mu_i + \sum_{x_j \in \mathbf{pa}_i^s} b_{ji} x_j, \sigma_i). \tag{2.1}$$

The local parameters θ_m^i are given by $(\mu_i, \sigma_i, \vec{b}_i)$.

2.2 Exponential Family of Distributions

The family of probability distributions having the form:

$$f(x|\theta) = a(\theta) b(x) e^{c(\theta) \cdot d(x)} \tag{2.2}$$

where $\theta \in \Theta$ and $c(\theta)$, $d(x)$ and θ are scalars or vectors is called the *exponential* family or sometimes the *Koopman-Darmois family* family. (e.g. [13] pages 362,370 and [12] page 161). This family can also be parameterized in the following form:

$$f(x|\theta) = e^{\hat{c}(\theta) \cdot \hat{d}(x) - \hat{a}(\theta)} \tag{2.3}$$

such that

$$\begin{aligned} \hat{c}(\theta) &= \langle c_1(\theta), \ldots, c_k(\theta), 1 \rangle \\ \hat{d}(x) &= \langle d_1(x), \ldots, d_k(x), \ln b(x) \rangle \\ \hat{a}(\theta) &= -\ln a(\theta) \end{aligned} \tag{2.4}$$

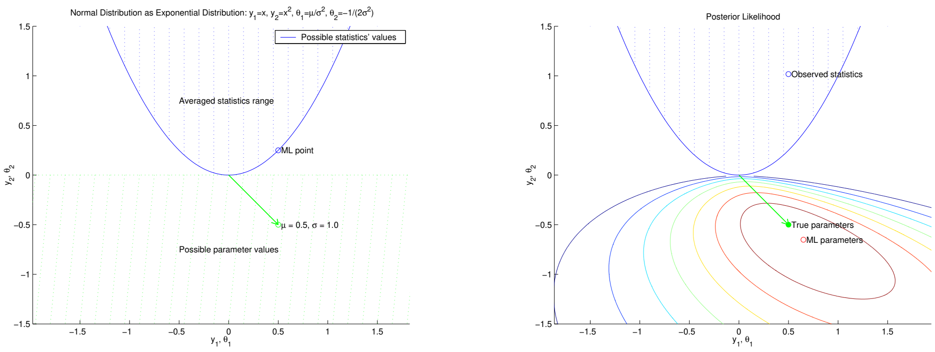
where $c_i(\theta)$ denotes the ith element of vector $c(\theta)$. Note that $\hat{c}$ may be considered as an arbitrary function of θ in parameterization 2.3, since the inverse parameterization from Equation 2.3 to Equation 2.2 is given by $c = \hat{c}$, $b(x) \equiv 1$ and $a(\theta) = e^{-\hat{a}(\theta)}$. It follows from [12] (page 156) that $\hat{d}(x)$ is a sufficient statistics for the given exponential family.

Since $\int f(x|\theta) dx = 1$, the value of $\hat{a}(\theta)$ is completely determined by the value of $\hat{c}(\theta)$. Thus we can, without loss-of-generality, use $Image(\hat{c})$ instead of Θ as a parameter space. In this way we get the following representation of exponential family, which was used by Schwarz [37]:

$$f(x|\theta) = e^{\theta \cdot y(x) - b(\theta)} \tag{2.5}$$

We adopt the exponential distribution as presented by Equation 2.5 and analyze its properties. The moment generating function ([12], page 26) of an exponential distribution is:

$$\begin{aligned} \psi(t) = E(e^{ty}) &= \int e^{ty} e^{\theta \cdot y - b(\theta)} dy \\ &= \int e^{(\theta + tI) \cdot y - b(\theta + tI)} e^{b(\theta + tI) - b(\theta)} dy = e^{b(\theta + tI) - b(\theta)} \end{aligned} \tag{2.6}$$



(a) (b)

Figure 2.1: Representation of normal distribution family as a subfamily of exponential family of distributions. (a) Graph shows the range of natural parameters $(-\infty, +\infty) \times (-\infty, 0)$; the range of possible statistics of each sample (parabola $y_2 = y_1^2$); the range of averaged statistics for sample from normal distribution (parabola interior); and specific parameter vector and point of maximum likelihood for this parameter vector. (b) Graph shows averaged statistics from sampling 100 points from distribution defined by parameters shown on (a). This graph also shows the posterior likelihood isolines and maximum likelihood point given the sampled data.

where I is a vector of ones of an appropriate dimension. Therefore

$$\bar{y} = E(y) = \psi'(0) = \nabla b(\theta) \tag{2.7}$$

and

$$Cov(y) = E\left((y - \bar{y})(y - \bar{y})^T\right) = E(yy^T) - \bar{y}\bar{y}^T = (\psi)^{(2)}(0) - \bar{y}\bar{y}^T = \mathcal{H}b(\theta) \tag{2.8}$$

where $\mathcal{H}$ denotes the Hessian operator, i.e. $\mathcal{H}b(\theta)$ is the matrix of second-order derivatives of $b(\cdot)$ evaluated at θ. Note, that $\mathcal{H}b(\theta)$ is positive definite, unless statistics y are linearly dependent ([12], page 25).

An analysis of asymptotic properties of parameter estimation for distributions from exponential families is performed in [26, 31, 4, 3]. This analysis is based on the application of differential geometry techniques [39, 40, 6] to the parametric space of exponential family.

An Example of Exponential Family - Normal Distribution

Consider a one-dimensional normal distribution, [12] page 37,

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma}\right]. \tag{2.9}$$

We can rewrite that in the exponential form:

$$\begin{aligned}
&f(x|\mu,\sigma^2) = e^{-b(\theta_1,\theta_2)+\sum_{i=1,2}\theta_i y_i(x)} \\
&y_1(x) = x, \quad y_2(x) = x^2, \\
&\theta_1 = \tfrac{\mu}{\sigma^2}, \quad \theta_2 = -\tfrac{1}{2\sigma^2} \\
&b(\theta_1,\theta_2) = -\tfrac{\theta_1^2}{4\theta_2} + \tfrac{1}{2}\ln\left[-\tfrac{\pi}{\theta_2}\right]
\end{aligned} \tag{2.10}$$

In this parameterization, the natural parameter space is $(-\infty,+\infty)\times(-\infty,0)$ and sufficient statistics y_1 and y_2 get the values from the parabola $y_2 = y_1^2$. This concept is illustrated on Figure 2.1. Note that the maximum likelihood value of $x=\mu$ corresponds to the point on the parabola where normal is parallel to the parameter vector.

Chapter 3

Bayesian Information Criterion

Schwarz proposed in his paper [37] an asymptotic criterion for model selection which is now termed the *Bayesian Information Criterion* (*BIC*). Given $X_1, \ldots, X_n$ *independent and identically distributed* (*i.i.d.*) observations and given a number of competing models the BIC procedure is to: *Choose the model for which* $\ln M_j(X_1, \ldots, X_n) - \frac{1}{2}k_j \ln n$ *is largest,* where k_j is dimension of the jth model, and M_j maximum of likelihood function for the jth model.

Schwarz proved the asymptotic correctness of this procedure for observations that come from a linear exponential family (LEF). Later this result has been extended by Haughton [20] for curved exponential families (CEF). This chapter derives in great detail the results in [37]. It is hoped that this careful derivation will enable us to extend Schwarz' results to stratified exponential models [18].

3.1 The exact Bayes procedure

The Bayesian model selection procedure chooses the model M of maximum posterior probability given the data D, which is

$$P(M|D) \propto P(M,D) = P(D|M)P(M) = P(M)\int P(D|M,\theta)P(\theta|M)d\theta \quad (3.1)$$

The asymptotic expansion of $\ln P(M|D)$ is an example of Laplace method for integrals (e.g [11, 32, 8, 22, 23]). This expansion is done under the following assumptions:

Assumption 1 (Exponential Distribution) *The observations* $X_1, \ldots, X_n$ *are i.i.d. and come from an exponential distribution:*

$$f(x|\theta) = \exp(\theta \cdot y(x) - b(\theta)). \quad (3.2)$$

where θ ranges over the natural parameter space Θ, which is a convex subset of $\mathbb{R}^K$.

Assumption 2 (Model linearity) *The models $\{m_j\}$ are represented by distinct affine subspaces of $\mathbb{R}^K$, i.e. $\forall\theta \in m_j, \exists\theta^j \in \mathbb{R}^{k_j}$, s.t. $\theta = A_j\theta^j + b_j$, where A_j is a $K \times k_j$ matrix and b_j is a K-dimensional vector defined by a model.*

After Schwarz [37], we use the differential geometry term *linear submanifold* [39, 40] to denote the affine subspace of $\mathbb{R}^K$.

Assumption 3 (Bayesian Prior) *The prior distribution is of the form:*

$$\mu = \sum \alpha_j \mu_j \tag{3.3}$$

where α is the a priori probability of the jth model being the true one, and μ_j is the conditional a priori distribution of θ given the jth model.

Assumption 4 (Boundedness) *The conditional prior distribution of θ given jth model, μ_j, has a k_j-dimensional density that is bounded and locally bounded away from zero throughout $m_j \cap \Theta$.*

Assumption 4 means that μ_j (a normalized measure induced on m_j by μ) is differentiable (i.e. $d\mu_j(\theta^j)$ exists for all $\theta_{k_j} \in \mathbb{R}^j$) and $\exists M, \forall\theta^j \in \mathbb{R}^{k_j}$, s.t. $\theta = A_j\theta^j + b_j \in \Theta$ holds $d\mu_j(\theta^j) < M$ (boundedness) and $\forall\theta^j \in \mathbb{R}^{k_j}, \exists U \subseteq \mathbb{R}^{k_j}$, U open, $\theta^j \in U$ and $\exists\epsilon > 0$, s.t. $\forall t \in U, d\mu_j(t) > \epsilon$ (local boundedness from zero).

Assumption 4 ensures that any submanifold of m_j which is of lower dimensionality than m_j will be of measure zero, since the intersection of two distinct linear manifolds either is one of them (impossible since that will contradict boundedness), or has a lower dimensionality than both.

Note that, unless there is only one model with $k = K$, the parameter prior μ is not absolutely continuous, since it puts positive probability on some lower-dimensional submanifolds of Θ, that correspond to the competing models.

Given Assumptions 1, 2 and 3, the Bayes solution for model selection (under fixed penalty for guessing the wrong model) consists of selecting the model that is a posteriori most probable, i.e. choose j that maximizes the posterior model probability. Taking logarithm of Equation 3.1, we get:

$$\ln P(M, D) = S(Y, n, j) = \ln \int \alpha_j \exp((Y \cdot \theta - b(\theta))n) d\mu_j(\theta), \tag{3.4}$$

where the integral extends over $m_j \cap \Theta$, and Y is the averaged sufficient statistics, $(1/n)\sum_{i=1}^n y(X_i)$.

The asymptotic evaluation of the integral presented by Equation 3.4 is done using the Assumptions 1-4 and the following assumption:

Assumption 5 (Internal Point) *For every m_j, the maximum of $Y \cdot \theta - b(\theta)$ for $\theta \in m_j \cap \Theta$ is achieved at some* internal *point of $m_j \cap \Theta$.*

3.2 Asymptotics

We are interested in asymptotic expansion of $S(y,n,j)$ as $n \to \infty$.

Proposition 1 (Schwarz' main result) *For fixed Y and j, as n tends to ∞,*

$$S(Y,n,j) = n\sup(Y\cdot\theta - b(\theta)) - \frac{1}{2}k_j \ln n + R \tag{3.5}$$

where the remainder $R = R(Y,n,j)$ is bounded in n for a fixed Y and j.

Note that in case we have a finite number of models, $R(Y,n,j)$ is bounded in n for fixed Y, and bound is independent of j.

The proof requires several lemmas:

Lemma 1 *The proposition holds when $Y\cdot\theta - b(\theta) = A - \lambda||\theta-\theta_0||^2$ where $\lambda > 0$, θ_0 is a fixed k_j-dimensional vector in m_j, and μ_j is Lebesgue measure on m_j.*

Note, that to formally prove this lemma we assume that m_j is a k_j dimensional affine subspace of $\mathbb{R}^K$ (Assumption 2).
Proof: Recall that for normal distribution (e.g. [12], page 51):

$$(2\pi)^{-k/2}|\Sigma|^{-\frac{1}{2}} \int_{\mathbb{R}^k} \exp\left[-\frac{1}{2}(x-\bar{x})^T\Sigma^{-1}(x-\bar{x})\right] dx = 1, \tag{3.6}$$

so the explicit evaluation of integral 3.4 yields

$$\int \alpha_j e^{(A-\lambda||\theta-\theta_0||^2)n} d\theta = \alpha_j(\pi/n\lambda)^{k_j/2}e^{nA}, \tag{3.7}$$

and

$$\sup_\theta\{A - \lambda||\theta-\theta_0||^2\} = A. \tag{3.8}$$

Therefore

$$S(Y,n,j) = \ln \alpha_j(\pi/n\lambda)^{k_j/2}e^{nA} = nA - \frac{1}{2}k_j\ln n + R \tag{3.9}$$

establishes the proposition in this case, with $R = \frac{1}{2}k_j \ln\frac{\pi}{\lambda} + \ln\alpha_j$. ■

Lemma 2 *If two bounded positive random variables U and V agree on the set where either exceeds ρ for some $0 < \rho < \operatorname{ess\,sup} U$, then*

$$\ln E(U^n) - \ln E(V^n) \to 0 \tag{3.10}$$

as $n \to \infty$.

Proof: We have two bounded positive random variables U and V such that for some ρ, $0 < \rho < \operatorname{ess\,sup} U$:

$$U = V \text{ for } V > \rho \text{ or } U > \rho \tag{3.11}$$

Lets first prove the lemma for V that vanishes where $U \leq \rho$, i.e. $V = 0$ for $U \leq \rho$. In this case $0 \leq U^n - V^n \leq \rho^n$, and therefore

$$E(V^n) \leq E(U^n) \leq E(V^n) + \rho^n = E(V^n)\left(1 + \frac{\rho^n}{E(V^n)}\right) \tag{3.12}$$

it is now sufficient to show that $\ln(1 + (\rho^n/E(V^n))) \to 0$ as $n \to 0$. Now $E^{1/n}(V^n) \to \operatorname{ess\,sup} V$ ([14], pages 78,87). and $\operatorname{ess\,sup} V = \operatorname{ess\,sup} U > \rho$ yields that $\rho/(E(V^n))^{1/n}$ a strictly less than 1 (beginning from some n_0), hence $\rho^n/E(V^n)$ tends to zero, and so does $\ln(1 + (\rho^n/E(V^n)))$.

Now, for a general V, define $\tilde{V}$ as:

$$\tilde{V} = \begin{cases} V & V > \rho \\ 0 & V \leq \rho \end{cases} \tag{3.13}$$

and similarly $\tilde{U}$. Since $\tilde{V} \leq V$ (and $\tilde{U} \leq U$) we have

$$-\left(\ln E(V^n) - \ln E(\tilde{U}^n)\right) \leq \ln E(U^n) - \ln E(V^n) \leq \ln E(U^n) - \ln E(\tilde{V}^n). \tag{3.14}$$

The right side and the left side of Equation 3.14 tend to zero as $n \to \infty$ (by the argument from previous paragraph) and thus $\ln[E(U^n) - \ln E(V^n)] \to 0$ as $n \to \infty$. ■

Lemma 3 *For some $0 < \rho < e^A$, where $A = \sup(Y \cdot \theta - b(\theta))$, a vector θ_0, and some positive λ_1 and λ_2, the following holds wherever $e^{Y \cdot \theta - b(\theta)} > \rho$:*

$$A - \lambda \|\theta - \theta_0\|^2 \leq Y \cdot \theta - b(\theta) \leq A - \lambda \|\theta - \theta_0\|^2. \tag{3.15}$$

Proof: The matrix of second order derivatives of $b(\theta)$ is the covariance matrix of y and under the assumption of y being linearly independent, $\mathcal{H}b(\theta)$ is positive definite for all θ of interest (see Section 2.2, and [12] page 25).

Therefore $Y \cdot \theta - b(\theta)$ is strictly convex, since $\mathcal{H}[Y \cdot \theta - b(\theta)] = -\mathcal{H}b(\theta)$ is negative definite and it attains maximum at θ_0, such that

$$\nabla [Y \cdot \theta - b(\theta)] = Y - \nabla b(\theta) = 0 \tag{3.16}$$

i.e. at θ_0 such that $\nabla b(\theta_0) = Y$. (Note, that under the Assumption 5, θ_0 is an internal point of $m_j \cap \Theta$.)

Consider a Taylor expansion of $\Phi(\theta) = Y \cdot \theta - b(\theta)$ around θ_0 ([27], page 173 and [28], page 349):

$$\Phi(\theta_0 + \Delta\theta) = \Phi(\theta_0) + (\Delta\theta \cdot \nabla)\Phi(\theta_0) + \frac{1}{2!}(\Delta\theta \cdot \nabla)^2 \Phi(\theta_0) + R_3 \tag{3.17}$$

where $(u \cdot \nabla)$ is a directional derivative operator, i.e. $(u \cdot \nabla)f = (\nabla f)^T u$, and $R_3 = O(||\Delta\theta||^3)$ is a remainder, which is bounded by $R_3 < c||\Delta\theta||^3$ on some sufficiently small neighborhood of θ_0. Taking the actual derivatives we have

$$\Phi(\theta_0+\Delta\theta) = Y^T\theta_0 - b(\theta_0) + (Y - \nabla b(\theta_0))^T\Delta\theta - \frac{1}{2}(\Delta\theta)^T \mathcal{H}b(\theta_0)(\Delta\theta) + R_3 \quad (3.18)$$

Since θ_0 is a maximum point, we have $Y - \nabla b(\theta_0) = 0$ and

$$\Phi(\theta_0 + \Delta\theta) = A - \frac{1}{2}(\Delta\theta)^T \mathcal{H}b(\theta_0)(\Delta\theta) + R_3 \quad (3.19)$$

where $A = sup(Y \cdot \theta - b(\theta))$. Let $\lambda'_1, \ldots, \lambda'_k$ be the eigenvalues of $\mathcal{H}b(\theta_0)$ and let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimal and maximal eigenvalue respectively. We have (e.g. [41, 34])

$$\frac{1}{2}\lambda_{\min}||\Delta\theta||^2 \leq \frac{1}{2}\Delta\theta^T \mathcal{H}b(\theta_0)\Delta\theta \leq \frac{1}{2}\lambda_{\max}||\Delta\theta||^2 \quad (3.20)$$

Let λ_1, λ_2 be a little bit larger and a little bit smaller than $\frac{1}{2}\lambda_{max}$ and $\frac{1}{2}\lambda_{\min}$ respectively, By strict convexity it is now easy to determine $\rho < e^A$ so that it will bound $e^{Y\cdot\theta - b(\theta)}$ outside the neighborhood $N(\theta_0)$ where R is dominated by $2\lambda_1 - \lambda_{\max}$ and by $\lambda_{\min} - 2\lambda_2$. I.e, for every $\theta = \theta_0 + \Delta\theta$ from $N(\theta_0)$

$$\begin{aligned} -\tfrac{1}{2}(\Delta\theta)^T \mathcal{H}b(\theta_0)(\Delta\theta) + R_3 \quad &\leq -\tfrac{1}{2}\lambda_{\min}||\Delta\theta||^2 + cr^3 \\ &\leq (-\tfrac{1}{2}\lambda_{\min} + cr)r^2 \leq -\lambda_2||\Delta\theta||^2 \end{aligned} \quad (3.21)$$

where $r = \max_{\theta\in N(\theta_0)} \Delta\theta$ and λ_2 is such that $\Delta\lambda_2 = \frac{1}{2}\lambda_{\min} - \lambda_2$ is larger than cr. Obtaining the lower bound similarly for λ_1 a little larger than $\frac{1}{2}\lambda_{\max}$, we have We have

$$A - \lambda_1||\Delta\theta||^2 \leq (Y \cdot \theta - b(\theta)) \leq A - \lambda_2||\Delta\theta||^2 \quad (3.22)$$

where $\theta = \theta_0 + \Delta\theta$ and $2\lambda_1$, $2\lambda_2$ are a little bit larger and a little bit smaller than all the eigenvalues of $\mathcal{H}b(\theta_0)$. ■

Note, that in order to prove correctness of Proposition 1, the neighborhood defined by ρ should be entirely inside $m_j \cap \Theta$.

Proof of the Proposition 1: For some specific m_j. Let

$$\begin{aligned} U(\theta) &= e^{A-\lambda_1||\theta-\theta_0||^2} \\ V(\theta) &= e^{Y\cdot\theta - b(\theta)} \\ W(\theta) &= e^{A-\lambda_2||\theta-\theta_0||^2} \end{aligned} \quad (3.23)$$

Let $N(\theta_0)$ denote the neighborhood of θ_0 where Lemma 2 holds, i.e. where $V(\theta)$ is greater than some $\rho \equiv \rho_V$. Let $\tilde{V}(\theta)$ equal to $V(\theta)$ in $N(\theta_0)$ and zero otherwise. Define $\tilde{U}$ and $\tilde{W}$ similarly. We can bound U, V and W outside that neighborhood, $N(\theta_0)$, by ρ_U, ρ_V and ρ_W and applying Lemma 2 we get

$$\begin{aligned} &S(Y,n,j) - \ln E(\tilde{V}^n) \to 0 \\ &\ln E(U^n) - \ln E(\tilde{U}^n) \to 0 \\ &\ln E(W^n) - \ln E(\tilde{W}^n) \to 0 \end{aligned} \tag{3.24}$$

From Lemma 3 the following holds on $N(\theta_0)$

$$e^{A-\lambda_1||\theta-\theta_0||^2} \leq e^{Y\cdot\theta - b(\theta)} \leq e^{A-\lambda_2||\theta-\theta_0||^2} \tag{3.25}$$

where $A = \sup(Y\cdot\theta - b(\theta))$. Thus for every θ, $\tilde{U}(\theta) \leq \tilde{V}(\theta) \leq \tilde{W}(\theta)$ and

$$E[\tilde{U}^n(\theta)] \leq E[\tilde{V}^n(\theta)] \leq E[\tilde{W}^n(\theta)] \tag{3.26}$$

We now evaluate $E[\tilde{U}^n(\theta)]$ and $E[\tilde{U}^n(\theta)]$, i.e. $\ln\int_{N(\theta_0)} e^{A-\lambda||\theta-\theta_0||^2}\mu_j(\theta)d\theta$ for $\lambda = \lambda_1, \lambda_2$. Let $c_1 = \min_{N(\theta_0)}\mu_j(\theta)$ and $c_2 = \max_{N(\theta_0)}\mu_j(\theta)$. By application of Lemma 1

$$\begin{aligned} E[\tilde{W}^n(\theta)] \quad &\leq \ln\textstyle\int_{N_\rho(\theta_0)} e^{A-\lambda_2||\theta-\theta_0||^2}\mu_j(\theta)d\theta \\ &\leq \ln c_2 + \ln\textstyle\int e^{A-\lambda_2||\theta-\theta_0||^2}d\theta = nA - \frac{1}{2}k_j\ln n + R_2 \end{aligned} \tag{3.27}$$

where $R_2 = \frac{1}{2}k_j\ln\frac{\pi}{\lambda_2} + \ln c_2$. Similarly from Lemma 1 and Lemma 2 we have

$$\begin{aligned} E[\tilde{U}^n(\theta)] \quad &\geq \ln\textstyle\int_{N_\rho(\theta_0)} e^{A-\lambda_1||\theta-\theta_0||^2}\mu_j(\theta)d\theta \\ &\geq \ln c_1 + \ln\textstyle\int_{N_\rho(\theta_0)} e^{A-\lambda_1||\theta-\theta_0||^2}d\theta \end{aligned} \tag{3.28}$$

and

$$\ln\int_{N_\rho(\theta_0)} e^{A-\lambda_1||\theta-\theta_0||^2}d\theta \to \ln\int e^{A-\lambda_1||\theta-\theta_0||^2}d\theta = nA - \frac{1}{2}k_j\ln n + \frac{1}{2}k_j\ln\frac{\pi}{\lambda_1} \tag{3.29}$$

so

$$E[\tilde{U}^n(\theta)] \to nA - \frac{1}{2}k_j\ln n + R_1 \tag{3.30}$$

where $R_1 = \frac{1}{2}k_j\ln\frac{\pi}{\lambda_1} + \ln c_1$. Combining Equations 3.24,3.26,3.27 and 3.30 the proposition is established with R bounded by $\frac{1}{2}k_j\ln\frac{\pi}{\lambda_2} + \ln c_2 + \ln\alpha_j$. ■

Note, that as $n\to\infty$, $\rho\to A$ (smaller $N(\theta_0)$'s) and $\lambda_1\to\lambda_{\max}$, $\lambda_2\to\lambda_{\min}$, the remainder R in approximation of $S(Y,n,j)$ is bounded by:

$$\frac{1}{2}k_j\ln\frac{\pi}{\lambda_{\max}} + \ln\mu(\theta_0) + \ln\alpha_j \leq R \leq \frac{1}{2}k_j\ln\frac{\pi}{\lambda_{\min}} + \ln\mu(\theta_0) + \ln\alpha_j \tag{3.31}$$

so the minimal error of approximation of $S(Y,n,j)$ dependents on the maximal eigenvalue of $\mathcal{H}b(\theta_0)$ (e.g. how sharp is the peak at θ_0) and the freedom in R

depends on the condition number of $\mathcal{H}b(\theta_0)$ (i.e. ratio between maximal and minimal eigenvalues).

In practical settings, we usually do not have a fixed Y, instead we have a sequence of $Y_1, Y_2, \ldots$ that is (hopefully) converging to Y. Haughton, [20], deals with that and other questions in her extension of the above result for curved exponential families.

Chapter 4

Parameter Priors for Discrete DAG Models

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 [9]. 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 [17, 21].

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 [16]. 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.

Our work (published as [35, 36]) gives a negative answer to the question of similarity between the discrete and continuous cases. In other words, it shows that, while global independence dictates a Normal-Wishart prior for Gaussian DAG models with more than 3 nodes, global 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.

4.1 Discrete DAG Models

We restrict our discussion to discrete DAG models, that describe a multinomial distribution on the discrete variables $X_1, \ldots, X_n$. In these models 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 (*p.d.f.*) on joint parameters and corresponding p.d.f.'s on network parameters for all complete models are everywhere positive and twice differentiable.*

This chapter 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 by Spiegelhalter and Lauritzen [38] to allow a decomposable prior-to-posterior analysis and global parameter independence for all the networks was introduced by Cooper and Herskovits [9] 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.

4.2 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 a complete two-node network, as shown on Figure 4.1, with variables X,Y having n and k states respectively. Since this network is complete it

Figure 4.1: A complete two node network for random variables X and Y.

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 $X \rightarrow Y$ (Figure 4.1) and $X \leftarrow Y$. Thus, according to the global independence assumption, such distributions should satisfy the following two functional equations, which encode global parameter independence:

$$\begin{array}{l} p(\{\theta_{ij}\}) = J_1^{-1} f_1(\{\theta_{i\cdot}\}) g_1(\{\theta_{j|i}\}) \\ p(\{\theta_{ij}\}) = J_2^{-1} f_2(\{\theta_{\cdot j}\}) g_2(\{\theta_{i|j}\}) \end{array} \tag{4.1}$$

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 [17] for two-node DAG models with binary variables.

Theorem 4 *Any probability distribution on $\{\theta_{ij}\}$ that satisfies the regularity assumption (1) and global parameter independence assumption (Equation 4.1), is of the form*

$$\begin{array}{l} p(\{\theta_{ij}\}) = C \left[\prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{\alpha_{ij}-1} \right] \cdot \\ \qquad H\left(\left\{ \frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,j}\theta_{i,j+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1 \right\} \right) \end{array} \tag{4.2}$$

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 4 implies that for two-node discrete DAG models global parameter independence assumption alone does not guarantee the Dirichlet distribution of priors. In Section 4.3 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 [17].

The proof of this theorem is based on the direct solution of a system of functional equations 4.1. The general approach is given in the following subsection.

4.2.1 The Functional Equation

In this subsection we present the functional equation that follows from Equation 4.1 and is the basis for the proof of Theorem 4.

Consider two sets of variables $\{y_i | 1 \le i \le n-1\}$ and $\{z_{ji} | 1 \le i \le n, 1 \le j \le k-1\}$. 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

$$\begin{array}{ll} y_n = 1 - \sum_{i=1}^{n-1} y_i & \\ z_{ki} = 1 - \sum_{j=1}^{k-1} z_{ji}, & 1 \le i \le n \\ x_j = \sum_{i=1}^{n} z_{ji} y_i, & 1 \le j \le k \\ w_{ji} = \frac{z_{ji} y_i}{x_j} & 1 \le j \le k,\ 1 \le i \le n. \end{array} \tag{4.3}$$

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

$$\begin{array}{ll} Y = (y_1, \ldots, y_{n-1}), & Z_i = (z_{1,i}, \ldots, z_{k-1,i}), \\ X = (x_1, \ldots, x_{k-1}), & W_j = (w_{j,1}, \ldots, w_{j,n-1}) \\ Z = (Z_1, \ldots, Z_n), & W = (W_1, \ldots, W_k) \end{array} \tag{4.4}$$

The functional equation we solve (4.1) can now be expressed as follows

$$F(Y)g(Z) = G(X)f(W) \tag{4.5}$$

by absorbing Jacobians appearing in Equation 4.1 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 4.5 are $y_1, \ldots, y_{n-1}$ and z_{ji}, $1 \le j \le k-1$, $1 \le i \le n$. All other variables appearing in Equation 4.5 are defined by Equations 4.3 and 4.4.

The solution of Equation 4.5 is based on the technique of reducing functional equations to partial differential equations ([1], page 324) and is presented in Appendix A.2. Similar technique was used in [17].

4.3 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:

$$p(\Theta_{\mathbf{X}}) = J_I^{-1} \prod_{j=1}^{n} f_{I,j}(\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}), \tag{4.6}$$

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 [21], Theorem 10).

4.3.1 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 4.6 we use Theorem 4. 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 5 *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 4. ■

Now, we apply Theorem 4 for $\mathbf{Y}_i$ and conclude that any $p(\Theta_{\mathbf{X}})$ that satisfies Equation 4.6 satisfies the following n equations (for $i = 1, \ldots, n$):

$$p(\Theta_{\mathbf{X}}) = C_i \prod_{r \in \mathbf{D}} \theta_r^{\alpha_{r,i}-1} H_i \left(\left\{ \frac{\theta_{r_i} \theta_{r_i'''}}{\theta_{r_i'} \theta_{r_i''}} \right\} \right) \tag{4.7}$$

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 4.2). I.e., when restricted to X_i: $[r_i]_{\perp X_i} = [r_i'']_{\perp X_i} \triangleq a$, $[r_i']_{\perp X_i} = [r_i''']_{\perp X_i} \triangleq b$ and $b = a+1$, and when restricted to $X \setminus X_i$: $[r_i]_{\perp X \setminus X_i} = [r_i']_{\perp X \setminus X_i} \triangleq c$, $[r_i'']_{\perp X \setminus X_i} = [r_i''']_{\perp X \setminus X_i} \triangleq d$ and $d = c+1$. Here $[r]_{\perp X}$ denote the vector of values of r for nodes $X \subseteq \mathbf{X}$. (We assume here the alphabetical order for the convenience, but the sets of parameters of H are equivalent under all orders.).

Lemma 5 specifies that the set of solutions of Equation 4.6 is equivalent to the solutions of Equation 4.7, which are obtained using the following lemma:

Lemma 6 *Consider the following system of m functional equations:*

$$\begin{array}{l} f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{1i}x_i + h_1(\vec{b}_{11}\vec{x},\ldots,\vec{b}_{1k_1}\vec{x}) \\ f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{2i}x_i + h_2(\vec{b}_{21}\vec{x},\ldots,\vec{b}_{2k_2}\vec{x}) \\ \vdots \\ f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{mi}x_i \\ \qquad\qquad +h_m(\vec{b}_{m1}\vec{x},\vec{b}_{m2}\vec{x},\ldots,\vec{b}_{mk_m}\vec{x}) \end{array} \tag{4.8}$$

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 4.8 is:

$$f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_i x_i + h(\vec{b}_1\vec{x},\ldots,\vec{b}_l\vec{x}) \tag{4.9}$$

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 4.7 can be transformed to the form of Equation 4.8 by taking a logarithm of both sides of each equation and changing the variables to $\ln\theta_r$, Lemma 6 provides a powerful tool for solving Equation 4.7. The proof of this lemma is quite straightforward by changing the variables inside the h-functions in such way that they include $\vec{b}_1\vec{x},\ldots,\vec{b}_l\vec{x}$. This proof is presented in Appendix A.1.

Application of Lemmas 5 and 6 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.

4.3.2 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 [17] 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, [16]).

Theorem 7 *Any distribution on $\Theta_{\mathbf{X}}$, where $\mathbf{X} = X_1,\ldots,X_n$ are binary random variables, that satisfies regularity assumption (1) and global parameter in-*

dependence assumption is of the form

$$p(\Theta_{\mathbf{X}}) = C \left[\prod_{\vec{x} \in \{0,1\}^n} \theta_{\vec{x}}^{\alpha_{\vec{x}}-1} \right] h \left(\frac{\prod_{\vec{x} \in A_0} \theta_{\vec{x}}}{\prod_{\vec{x} \in A_1} \theta_{\vec{x}}} \right) \tag{4.10}$$

where h is an arbitrary Lebesgue integrable, twice-differentiable and everywhere positive 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 proof, based on Lemmas 5 and 6, is explicated in the Appendix A.3.

4.4 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 8 *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 work.

Theorem 9 *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.*

$$p(\theta_{m_1}|m_1) = p(\theta^{i_1}_{m_1}|m_1)p(\theta_{m_1} \setminus \theta^{i_1}_{m_1}|m_1) \tag{4.11}$$

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

$$\begin{array}{l} p(\theta_{m_2}|m_2) = \\ \quad p(\theta_{m_2} \setminus \theta_{m_2}^{j_n}|m_2) \prod_{\mathbf{pa}_{j_n}^{s_2} \in \mathbf{D}_{\mathbf{Pa}_{j_n}^{s_2}}} p(\theta_{X_{j_n}|\mathbf{pa}_{j_n}^{s_2}}|m_2) \end{array} \tag{4.12}$$

where $\theta_{X_i|\mathbf{pa}_i^s} = \{\theta_{x_i|\mathbf{pa}_i^s}|x_i \in D_i\}$, *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 4.11 or Equation 4.12. Since Lemma 8 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 4.11 and 4.12 transform to:

$$\begin{array}{l} p(\{\theta_{ij}\}) = f_1(\{\theta_{i\cdot}\}_{i=1}^{n-1}) g_1(\{\theta_{j|i}\}_{i=1,\ldots,n}^{j=1,\ldots,k-1}) \\ p(\{\theta_{ij}\}) = f_2(\{\theta_{\cdot j}\}_{j=1}^{k-1}) \prod_{j=1}^{k} h_j(\{\theta_{i|j}\}_{i=1}^{n-1}) \end{array} \tag{4.13}$$

Any solution p that satisfies Equation 4.13 satisfies also Equation 4.1 and thus can be written in form given by Theorem 4 (Equation 4.2). We have

$$\begin{array}{l} C\left[\prod_{i=1}^{n}\prod_{j=1}^{k}\theta_{ij}^{\alpha_{ij}}\right] H\left(\left\{\frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,i}\theta_{i,j+1}}|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) \\ \quad = f_2(\{\theta_{\cdot j}\}_{j=1,\ldots,k-1}) \prod_{j=1}^{k} h_j(\{\theta_{i|j}\}_{i=1,\ldots,n-1}) \end{array} \tag{4.14}$$

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 $\theta_{\cdot j}$):

$$\hat{H}\left(\left\{\frac{w_{ji}w_{j+1,i+1}}{w_{j+1,i}w_{j,i+1}}|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) = \sum_{j=1}^{k} \hat{h}_j(\{w_{ji}\}_{i=1}^{n-1}) \tag{4.15}$$

the solution of which is

$$h_j(w_{j1},\ldots,w_{j,n-1}) = \beta_j \prod_{i=1}^{n} w_{ji}^{\beta_{ji}}, \quad 1 \le j \le k \tag{4.16}$$

where $\beta_j, \beta_{j1}, \ldots, \beta_{j,n}$ are constants. Combining the results of solution of Equation 4.13 for f_2 and Equation 4.16 we conclude the proof. ■

4.5 Discussion

This work shows that local parameter independence is essential in the characterization of a Dirichlet prior via discrete DAG models (Section 4.4, Theorem 9). In addition, the functional form of prior distributions that arise from global parameter independence assumption alone are investigated (Sections 4.2 and 4.3, Theorem 7).

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 5 and 6) to specify such distribution (in symbolic form) for any specific DAG model (see [35]).

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 9 (Equation 4.13) using the techniques presented in [24], as done in [17, 16].

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 the second condition in Theorem 9 is essential, or either it is enough to assume local independence alone. The integral functional equation that arises from this reduced set of assumptions is of the form (for two binary variables):

$$\begin{aligned} g_0(z_0)g_1(z_1) = \int_0^1 & G\left(z_0 y + z_1(1-y)\right) \\ & f\left(\frac{z_0 y}{z_0 y + z_1(1-y)}, \frac{(1-z_0)y}{1-z_0 y - z_1(1-y)}\right) dy \end{aligned} \tag{4.17}$$

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.

Chapter 5

Future Work

A number of theoretical and practical issues in model selection for Bayesian networks will be explored. Among them problems in the field of model selection for DAG models with hidden variables (Chapters 2 and 3):

- Classification of singularities that arise in the parametric representation of DAG models with hidden variables as a subfamily of the exponential family of distributions [18].
- Analytic computation of the effective dimensionality of parameters for Bayesian network with hidden nodes [19].
- Development of Bayesian Information Criterion for model selection for Bayesian networks with hidden nodes [18].

Additional problems that need further attention arise in research of functional form of prior distributions under various independence assumptions (Section 4.5):

- Remove the regularity assumption in the derivations of functional forms of prior distributions for discrete DAG models.
- Investigate the functional form of prior distribution that arise under local independence assumption alone.

In the applications area we will pick up the following task:

- To apply and evaluate methods for model selection for Bayesian networks in the field of genetic linkage analysis [29].

Bibliography

[1] Janos Aczél. *Lectures on Functional Equations and Their Applications*. Number 19 in Mathematics in Science and Engineering. Academic Press, 1966.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[3] O. E. Barndorff-Nielsen. *Information and Exponential Families In Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1978.

[4] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1994.

[5] Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses: Theory and Applications*. John Wiley & Sons, 1974.

[6] Marcel Berger and Bernard Gostiaux. *Differential Geometry: Manifolds, Curves, and Surfaces*. Number 115 in Graduate Texts in Mathematics. Springer-Verlag, 1988.

[7] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley series in probability and mathematical statistics. John Wiley & Sons, 1994.

[8] Norman Bleistein and Richard A. Hendelsman. *Asymptotic Expansions of Integrals*. Holt, Rinehart and Winston, 1975.

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

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

[11] N. G. De Bruijn. *Asymptotic Methods in Analysis*, volume 4 of *A Series of Monographs on Pure and Applied Mathematics*. North-Holland Publishing Company, 3rd edition, 1970.

[12] Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, 1970.

[13] Morris H. DeGroot. *Probability and Statistics*. Addison-Wesley, 2nd edition, 1986.

[14] Joseph L. Doob. *Measure Theory*. Number 143 in Graduate Texts in Mathematics. Springer-Verlag, 1994.

[15] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

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

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

[18] Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. To appear in Annals of Statistics.

[19] Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. In E. Horvitz and F. Jensen, editors, *Uncertainty in Artificial Intelligence*, volume 12, pages 283–290. Morgan Kaufmann, August 1996.

[20] 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.

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

[22] L. C. Hsu. A theorem on the asymptotic behavior of a multiple integral. *Duke Mathematical Journal*, pages 623–632, 1948.

[23] L. C. Hsu. On the asymptotic evaluation of a class of multiple integrals involving a parameter. *American Journal of Mathematics*, 73(3):625–634, July 1951.

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

[25] Finn V. Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[26] Robert E. Kass and Paul W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1997.

[27] Granio A. Korn and Theresa M. Korn. *"Spravochnik po Mathematike" a translation of "Mathematical Handbook"*. Nauka, 1974.

[28] Serge Lang. *Real and Functional Analysis*. Number 142 in Graduate Texts in Mathematics. Springer-Verlag, 3rd edition, 1993.

[29] Kenneth Lange. *Mathematical and Statistical Methods for Genetic Analysis*. Statistics for Biology and Health. Springer, 1997.

[30] Steffen L. Lauritzen. *Graphical Models*. Number 17 in Oxford Statistical Science Series. Clarendon Press, 1996.

[31] Michael K. Murray and John W. Rice. *Differential Geometry and Statistics*. Number 48 in Monographs and Statistics and Applied Probability. Chapman and Hall, 1993.

[32] Frank W. J. Olver. *Asymptotics and Special Functions*. Computer Science and Applied Mathematics: A Series of Monographs and Textbooks. Academic Press, 1974.

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

[34] Steven Roman. *Advanced Linear Algebra*. Number 135 in Graduate texts in mathematics. Springer-Verlag, 1992.

[35] Dmitry Rusakov and Dan Geiger. On parameter priors for discrete dag models. Technical Report CIS-2000-08, Technion - Israel Institute of Technology, 2000.

[36] Dmitry Rusakov and Dan Geiger. On parameter priors for discrete dag models. To Appear in Proceedings of AISTATS-2001, January 2001.

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

[38] David J. Spiegelhalter and Steffen L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

[39] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, Inc., 2nd edition, 1979.

[40] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 2. Publish or Perish, Inc., 2nd edition, 1979.

[41] Riaz A. Usmani. *Applied Linear Algebra*. Number 105 in Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, 1987.

[42] Sumio Watanabe. Algebraic analysis for non-regular learning machines. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 conference*, pages 356–362. MIT Press, 2000.

Appendix A

Proofs and Derivations

A.1 Useful Lemmas

Here, we prove two lemmas that are useful for solving functional equations presented in this work.

We first prove Lemma 6 from Section 4.3.1.

Lemma 6 *Consider the following system of m functional equations:*

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

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 A.1 is:

$$f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\alpha_i x_i+h(\vec{b}_1\vec{x},\ldots,\vec{b}_l\vec{x}) \tag{A.2}$$

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 $\vec{b}_{i1},\ldots,\vec{b}_{ik_i}$.

Proof: The proof is given for $m = 2$, but can be easily extended for a general case by induction. Suppose that we have a following system of functional equations:

$$\begin{array}{l} f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\alpha_i x_i+h(\vec{a}_1^T\vec{x},\ldots,\vec{a}_k^T\vec{x}) \\ f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\beta_i x_i+g(\vec{b}_1^T\vec{x},\ldots,\vec{b}_{k'}^T\vec{x}) \end{array} \tag{A.3}$$

By comparing the two equations we have:

$$h(\vec{a}_1^T\vec{x},\ldots,\vec{a}_k^T\vec{x}) = \sum_{i=1}^{n}(\beta_i-\alpha_i)x_i + g(\vec{b}_1^T\vec{x},\ldots,\vec{b}_{k'}^T\vec{x}) \tag{A.4}$$

It is easy to see that $V = \{\vec{a}^T\vec{x}|\vec{a}\in\mathbb{R}^n, \vec{x}=(x_1,\ldots,x_n)^T\}$ is a vector space and that $V_h = span(\vec{a}_1^T\vec{x},\ldots,\vec{a}_k^T\vec{x})$ and $V_g = span(\vec{b}_1^T\vec{x},\ldots,\vec{b}_{k'}^T\vec{x})$ are subspaces of V. We can now change the variables $x_1,\ldots,x_n$ to n-independent variables $y_1,\ldots,y_n$ where $y_i=\vec{\gamma}_i^T\vec{x}$ such that:

$$\begin{array}{l} V_h\bigcap V_g = span(y_1,\ldots,y_{k^*})\\ V_h = span(y_1,\ldots,y_{k^*},y_{k^*+1},\ldots,y_k)\\ V_g = span(y_1,\ldots,y_{k^*},y_{k+1},\ldots,y_{k+k'-k^*})\\ V = span(y_1,\ldots,y_n)\end{array} \tag{A.5}$$

where $k* = dim(V_h\bigcap V_g)$. Let Γ_h be the $n\times k$ matrix with columns $\vec{\gamma}_1,\ldots,\vec{\gamma}_k$, let Γ_g be the $n\times k'$ matrix with columns $\vec{\gamma}_1,\ldots,\vec{\gamma}_{k^*},\vec{\gamma}_{k+1},\ldots,\vec{\gamma}_{k+k'-k^*}$ and let Γ denote the $n\times n$ matrix with columns $\vec{\gamma}_1,\ldots,\vec{\gamma}_n$. Equation A.4 now transforms to:

$$\hat{h}(y_1,\ldots,y_{k^*},y_{k^*+1},\ldots,y_k) = \sum_{i=1}^{n}\delta_i y_i + \hat{g}(y_1,\ldots,y_{k^*},y_{k+1},\ldots,y_{k+k'-k^*}) \tag{A.6}$$

where

$$\begin{array}{l}\hat{h}(t_1,\ldots,t_k) = h((\Gamma_h^+\vec{a}_1)^T\vec{t},\ldots,(\Gamma_h^+\vec{a}_k)^T\vec{t})\\ \hat{g}(t_1,\ldots,t_{k'}) = g((\Gamma_g^+\vec{b}_1)^T\vec{t},\ldots,(\Gamma_g^+\vec{b}_{k'})^T\vec{t})\\ \vec{\delta} = \Gamma^{-1}(\vec{\beta}-\vec{\alpha})\end{array} \tag{A.7}$$

and $\Gamma_h^+\vec{a}_i$ (or $\Gamma_g^+\vec{b}_i$) are coordinates of $\vec{a}_i^T\vec{x}$ (or $\vec{b}_i^T\vec{x}$) in the corresponding basis of y's. For any matrix B, B^+ denotes pseudo-inverse, $B^+=(B^TB)^{-1}B^T$, which is well defined since Γ_h, Γ_g are of full rank and $n>k,k'$ (since $rank(B^TB) = rank(B)$, see [5] page 20).

Equation A.6 can be further transformed to:

$$\begin{array}{l}\tilde{h}(y_1,\ldots,y_{k^*},y_{k^*+1},\ldots,y_k)\\ \quad = \sum_{i=k+k'-k^*+1}^{n}\delta_i y_i + \tilde{g}(y_1,\ldots,y_{k^*},y_{k+1},\ldots,y_{k+k'-k^*})\end{array} \tag{A.8}$$

where

$$\begin{array}{l}\tilde{h}(t_1,\ldots,t_k) = \hat{h}(t_1,\ldots,t_k) - \sum_{i=1}^{k}\delta_i t_i\\ \tilde{g}(t_1,\ldots,t_{k'}) = \hat{g}(t_1,\ldots,t_{k'}) + \sum_{i=1}^{k'-k^*}\delta_{k+i}t_{k^*+i}\end{array} \tag{A.9}$$

Now it is clear from Equation A.8 that $\delta_i = 0$ for $i = k+k'-k^*+1,\ldots,n$ and functions $\tilde{h}$ and $\tilde{g}$ depend only on variables $y_1,\ldots,y_{k^*}$. Going back to original variables $x_1,\ldots,x_n$ we can obtain the solution to the original system A.3, which is:

$$f(x_1,\ldots,x_n) = \sum_{i=1}^{n}\lambda_i x_1 + \bar{h}(\vec{\gamma}_1^T\vec{x},\ldots,\vec{\gamma}_{k^*}^T\vec{x}) \tag{A.10}$$

where $\{\lambda_i\}_{i=1}^n$ are arbitrary constants, $\bar{h}$ is an arbitrary function of k^* variables and $\vec{\gamma}_1^T\vec{x}, \ldots, \vec{\gamma}_{k^*}^T\vec{x}$ is the basis for $V_h \bigcap V_g$. ■

Lemma 10 *Let $\mathbf{t} = \langle t_1, \ldots, t_n\rangle$ be n dimensional vector of variables (or values) and let $\tilde{\mathbf{t}}$ denote the first $n-1$ elements of $\mathbf{t}$, i.e. $\tilde{\mathbf{t}} = \langle t_1, \ldots, t_{n-1}\rangle$. Let $\Delta_m = \{\langle\alpha_1, \ldots, \alpha_m\rangle | \alpha_i > 0, \sum_{i=1}^m \alpha_i < 1\}$ denote the m dimensional unit simplex.*

Consider the following general type of system of functional equations:

$$f(t_1, \ldots, t_n) = h_i(\mathbf{t}, f_i(g_i(\mathbf{t}))),\ i = 1, \ldots, k \tag{A.11}$$

where $f, f_1, \ldots, f_k$ are unknown functions and $\mathbf{t} = \langle t_1, \ldots, t_n\rangle \in D \subseteq \mathbb{R}^n$ are independent variables. Suppose D is such that $\{\mathbf{t}|\mathbf{t} > 0, \sum_{j=1}^n t_j = 1\} \subset D$.

Consider the same equation, with $t_1, \ldots, t_n$ being dependent, with $t_n = 1 - \sum_{j=1}^{n-1} t_j$:

$$\tilde{f}(\tilde{\mathbf{t}}) = h_i(q(\tilde{\mathbf{t}}), f_i(g_i(q(\tilde{\mathbf{t}})))),\ i = 1, \ldots, k \tag{A.12}$$

where $q(\tilde{\mathbf{t}}) = \langle t_1, \ldots, t_{n-1}, 1 - \sum_{j=1}^{n-1} t_j\rangle$ and $\tilde{\mathbf{t}} \in \Delta_{n-1}$.

The relationship between the solutions of Equations A.11 and A.12 is:

1. *For any solution f of Equation A.11, $\tilde{f}(\tilde{\mathbf{t}}) \triangleq f(q(\tilde{\mathbf{t}}))$ is a solution for Equation A.12.*

2. *If h_i and g_i are scale-independent on D, i.e. $h_i(a\mathbf{t}, b) = h_i(\mathbf{t}, b)$ and $g_i(a\mathbf{t}) = g_i(\mathbf{t})$ for all $i = 1, \ldots, k$, $a, b \in \mathbb{R}$ ($a \neq 0$) and $\mathbf{t}, a\mathbf{t} \in D$. Then any solution $\tilde{f}$ of Equation A.12 is of the form $\tilde{f}(\tilde{\mathbf{t}}) \triangleq f(q(\tilde{\mathbf{t}}))$, where f is a solution for Equation A.11.*

Proof: The first part: Since f is a solution for Equation A.11 there exist functions $f_1, \ldots, f_k$ such that Equation A.11 holds for all $\mathbf{t} \in D$. We have (for $1 \leq i \leq k$):

$$\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}})) = h_i(q(\tilde{\mathbf{t}}), f_i(g_i(q(\tilde{\mathbf{t}})))) \tag{A.13}$$

for all $\tilde{\mathbf{t}} \in \Delta_{n-1}$. Thus $\tilde{f}$ is a solution of Equation A.12.

The second part: We first show that for any $\tilde{f}$ a solution for Equation A.12 $f(\mathbf{t}) \triangleq \tilde{f}(\tilde{\mathbf{t}}/\sigma)$ is a solution for Equation A.11, where $\sigma = \sum_{j=1}^n t_j$. Indeed, since $\tilde{f}$ is a solution for Equation A.12 there exist functions $f_1, \ldots, f_k$ such that Equation A.12 holds for all $\tilde{\mathbf{t}} \in \Delta_{n-1}$ (and for all $\mathbf{t} \in D$, $\tilde{\mathbf{t}}/\sigma \in \Delta_{n-1}$) We have (for $1 \leq i \leq k$):

$$\begin{aligned} f(\mathbf{t}) = \tilde{f}(\tilde{\mathbf{t}}/\sigma) &= h_i(q(\tilde{\mathbf{t}}/\sigma), f_i(g_i(q(\tilde{\mathbf{t}}/\sigma)))) \\ &= h_i(\mathbf{t}/\sigma, f_i(g_i(\mathbf{t}/\sigma))) = h_i(\mathbf{t}, f_i(g_i(\mathbf{t}))) \end{aligned} \tag{A.14}$$

for all $\mathbf{t} \in D$.

Now it is easy to check that $\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}}))$. We have:

$$f(q(\tilde{\mathbf{t}})) = \tilde{f}(\tilde{\mathbf{t}}/1) = \tilde{f}(\tilde{\mathbf{t}}). \tag{A.15}$$

Thus any solution for Equation A.12 can be found by using a solution of Equation A.11. ■

A.2 The Solution of Functional Equation for Two Node Networks

We now solve Equation 4.5 for any n, k. We use the following notations. Let $\hat{h}(x)$ denote $\ln h(x)$ for any positive function h. Also let

$$\begin{array}{ll}
\hat{F}_i(Y) = \frac{\partial \hat{F}(Y)}{\partial y_i} & 1 \le i \le n-1 \\
\hat{F}_{i_1 i_2}(Y) = \frac{\partial^2 \hat{F}(Y)}{\partial y_{i_1} \partial y_{i_2}} & 1 \le i_1, i_2 \le n-1 \\
\hat{g}_{ji}(Z) = \frac{\partial \hat{g}(Z)}{\partial z_{ji}} & 1 \le i \le n,\ 1 \le j \le k-1 \\
\hat{g}_{(j_1 i_1)(j_2 i_2)}(Z) = \frac{\partial^2 \hat{g}(Z)}{\partial z_{j_1 i_1} \partial z_{j_2 i_2}} & 1 \le i_1, i_2 \le n,\ 1 \le j_1, j_2 \le k-1
\end{array} \tag{A.16}$$

and similarly for G and f.

Taking partial derivatives of X and W gives

$$\begin{array}{ll}
\frac{\partial x_j}{\partial y_i} = z_{ji} - z_{jn} & \begin{array}{l} 1 \le j \le k, \\ 1 \le i \le n-1 \end{array} \\
\frac{\partial x_m}{\partial z_{ji}} = \left\{ \begin{array}{ll} y_i & j = m \\ 0 & j \ne m \end{array} \right. , \quad \frac{\partial x_k}{\partial z_{ji}} = -y_i & \begin{array}{l} 1 \le m, j \le k-1, \\ 1 \le i \le n \end{array} \\
\frac{\partial w_{ji}}{\partial y_l} = \frac{\partial \frac{z_{ji} y_i}{x_j}}{\partial y_l} = \left\{ \begin{array}{ll} -\frac{z_{ji} y_i (z_{jl} - z_{jn})}{x_j^2} & i \ne l \\ -\frac{z_{ji} y_i (z_{jl} - z_{jn})}{x_j^2} + \frac{z_{ji}}{x_j} & i = l \end{array} \right. & \begin{array}{l} 1 \le j \le k, \\ 1 \le i, l \le n-1 \end{array} \\
\frac{\partial w_{ji}}{\partial z_{ml}} = \frac{\partial \frac{z_{ji} y_i}{x_j}}{\partial z_{ml}} = \left\{ \begin{array}{ll} -\frac{z_{ji} y_i y_l}{x_j^2} & i \ne l, m = j \\ -\frac{z_{ji} y_i y_l}{x_j^2} + \frac{y_i}{x_j} & i = l, m = j \\ 0 & m \ne j \end{array} \right. & \begin{array}{l} 1 \le j, m \le k-1, \\ 1 \le i, l \le n,\ i \ne n \end{array} \\
\frac{\partial w_{ki}}{\partial z_{ml}} = \frac{\partial \frac{z_{ki} y_i}{x_k}}{\partial z_{ml}} = \left\{ \begin{array}{ll} \frac{z_{ki} y_i y_l}{x_k^2} & i \ne l \\ \frac{z_{ki} y_i y_l}{x_k^2} - \frac{y_i}{x_k} & i = l \end{array} \right. & \begin{array}{l} 1 \le m \le k-1, \\ 1 \le i, l \le n,\ i \ne n \end{array}
\end{array} \tag{A.17}$$

(Note, that there is no error in line 3: $\frac{\partial \frac{z_{ki} y_i}{x_k}}{\partial y_l}$ indeed can be written in the given form). Additional calculations give:

$$\begin{array}{ll}
\frac{\partial \hat{G}}{\partial y_i} = \sum_{j=1}^{k-1} \left[(z_{ji} - z_{jn}) \hat{G}_j \right] & 1 \le i \le n-1 \\
\frac{\partial \hat{G}}{\partial z_{ji}} = y_i \hat{G}_j & \begin{array}{l} 1 \le i \le n, \\ 1 \le j \le k-1 \end{array} \\
\frac{\partial \hat{f}}{\partial y_i} = \sum_{j=1}^{k} \left(\sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l (z_{ji} - z_{jn})}{x_j^2} \hat{f}_{jl} \right] + \frac{z_{ji}}{x_j} \hat{f}_{ji} \right) & 1 \le i \le n-1 \\
\frac{\partial \hat{f}}{\partial z_{ji}} = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l y_i}{x_j^2} \hat{f}_{jl} + \frac{z_{kl} y_l y_i}{x_k^2} \hat{f}_{kl} \right] + \frac{y_i}{x_j} \hat{f}_{ji} - \frac{y_i}{x_k} \hat{f}_{ki} & \begin{array}{l} 1 \le i \le n-1, \\ 1 \le j \le k-1 \end{array} \\
\frac{\partial \hat{f}}{\partial z_{jn}} = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l y_n}{x_j^2} \hat{f}_{jl} + \frac{z_{kl} y_l y_n}{x_k^2} \hat{f}_{kl} \right] & 1 \le j \le k-1
\end{array} \tag{A.18}$$

Let $\hat{C}_j = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l}{x_j^2} \hat{f}_{jl} \right]$. By taking the logarithm and then a derivative wrt

y_i $(1 \leq i \leq n-1)$ of Equation 4.5, we get,

$$\hat{F}_i(Y) = \sum_{j=1}^{k-1} \left[(z_{ji} - z_{jn}) \hat{G}_j(X) \right] + \sum_{j=1}^{k} \left[(z_{ji} - z_{jn}) \hat{C}_j(x_j, W) + \frac{z_{ji}}{x_j} \hat{f}_{ji}(W) \right] \tag{A.19}$$

By taking the logarithm and then a derivative wrt z_{ji} $(1 \leq i \leq n-1, 1 \leq j \leq k-1)$ of Equation 4.5, we get,

$$\hat{g}_{ji}(Z) = y_i \hat{G}_j(X) + y_i \hat{C}_j(x_j, W) - y_i \hat{C}_k(x_k, W) + \frac{y_i}{x_j} \hat{f}_{ji}(W) - \frac{y_i}{x_k} \hat{f}_{ki}(W) \tag{A.20}$$

and by the same operation wrt z_{jn} $(1 \leq j \leq k-1)$:

$$\hat{g}_{jn}(Z) = y_n \hat{G}_j(X) + y_n \hat{C}_j(x_j, W) - y_n \hat{C}_k(x_k, W) \tag{A.21}$$

From Equations A.20 and A.21 we get: $(1 \leq i \leq n-1, 1 \leq j \leq k-1)$

$$\frac{1}{y_i} \hat{g}_{ji}(Z) - \frac{1}{y_n} \hat{g}_{jn}(Z) = \frac{1}{x_j} \hat{f}_{ji}(W) - \frac{1}{x_k} \hat{f}_{ki}(W) \tag{A.22}$$

Solving Equation A.22 for $\frac{1}{x_j} \hat{f}_{ji}(W)$ and substitution into Equation A.21 gives:

$$\sum_{l=1}^{n} \frac{z_{jl}}{x_j} \hat{g}_{jl}(Z) = \hat{G}_j(X) - \sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \frac{1}{x_k} \hat{f}_{kl}(W) \tag{A.23}$$

Simplifying Equation A.19 by using Equations A.21,A.22 and recalling $z_{ki} = 1 - \sum_{j=1}^{k-1} z_{ji}$ we get (for $1 \leq i \leq n-1$):

$$\hat{F}_i(Y) = \sum_{j=1}^{k-1} \left(\frac{z_{ji}}{y_i} \hat{g}_{ji}(Z) - \frac{z_{jn}}{y_n} \hat{g}_{jn}(Z) \right) + \frac{1}{x_k} \hat{f}_{ki}(W) \tag{A.24}$$

Multiplying Equations A.24 by $(w_{ji} - w_{ki})$, taking the sum of the resulting equations and substitution $\sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \frac{1}{x_k} \hat{f}_{kl}(W)$ from Equation A.23 gives (for $1 \leq j \leq k-1$):

$$\begin{aligned} \textstyle\sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \hat{F}_l(Y) = & \\ & \textstyle\sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \left[\sum_{m=1}^{k-1} \left(\frac{z_{ml}}{y_l} \hat{g}_{ml}(Z) - \frac{z_{mn}}{y_n} \hat{g}_{mn}(Z) \right) \right] \\ & + \hat{G}_j(X) - \textstyle\sum_{l=1}^{n} \frac{z_{jl}}{x_j} \hat{g}_{jl}(Z) \end{aligned} \tag{A.25}$$

After some simplifications (recall that $w_{ji} = \frac{z_{ji} y_i}{x_j}$) we get:

$$\begin{aligned} \textstyle\sum_{l=1}^{n-1} (w_{jl} - w_{kl}) \hat{F}_l(Y) = & \textstyle\sum_{l=1}^{n} \left[\left(\frac{z_{jl}}{x_j} - \frac{z_{kl}}{x_k} \right) \sum_{m=1}^{k-1} z_{ml} \hat{g}_{ml}(Z) \right] \\ & + \hat{G}_j(X) - \textstyle\sum_{l=1}^{n} \frac{z_{jl}}{x_j} \hat{g}_{jl}(Z) \end{aligned} \tag{A.26}$$

Taking a derivative by z_{ji} we get:

$$\sum_{l=1}^{n-1}\left(-\frac{z_{jl}y_l y_i}{x_j^2}-\frac{z_{kl}y_l y_i}{x_k^2}\right)\hat{F}_l(Y)+\frac{y_i}{x_j}\hat{F}_i(Y)+\frac{y_i}{x_k}\hat{F}_i(Y)= \tag{A.27}$$
$$\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{x_j}-\frac{z_{kl}}{x_k}\right)\sum_{m=1}^{k-1}z_{ml}\hat{g}_{(ml)(ji)}(Z)\right]+\left(\frac{z_{ji}}{x_j}-\frac{z_{ki}}{x_k}\right)\hat{g}_{ji}(Z)$$
$$+\sum_{l=1}^{n}\left[\left(-\frac{z_{jl}y_i}{x_j^2}-\frac{z_{kl}y_i}{x_k^2}\right)\sum_{m=1}^{k-1}z_{ml}\hat{g}_{ml}(Z)\right]+\left(\frac{1}{x_j}+\frac{1}{x_k}\right)\sum_{m=1}^{k-1}z_{mi}\hat{g}_{mi}(Z)$$
$$+y_i\hat{G}_{jj}(X)-\sum_{l=1}^{n}\frac{z_{jl}}{x_j}\hat{g}_{(jl)(ji)}(Z)+\sum_{l=1}^{n}\frac{z_{jl}y_i}{x_j^2}\hat{g}_{jl}(Z)-\frac{1}{x_j}\hat{g}_{ji}(Z)$$

Substituting $z_{ji}=\frac{1}{k}$ (and thus $x_j=\frac{1}{k}$, $w_{ji}=y_i$) for $1\le i\le n$, $1\le j\le k$ we get ($1\le i\le n-1$):

$$-\sum_{l=1}^{n-1}y_l\hat{F}_l(Y)+\hat{F}_i(Y)=\frac{1}{y_i}C_i+A \tag{A.28}$$

where

$$C_i=\frac{1}{k}\sum_{m=1}^{k-1}\hat{g}_{mi}(\frac{1}{k})-\frac{1}{2k}\sum_{l=1}^{n}\hat{g}_{(jl)(ji)}(\frac{1}{k})-\frac{1}{2}\hat{g}_{ji}(\frac{1}{k}) \tag{A.29}$$
$$A=-\sum_{l=1}^{n}\sum_{m=1}^{k-1}\hat{g}_{ml}(\frac{1}{k})+\frac{1}{2k}\hat{G}_{jj}(\frac{1}{k})+\frac{1}{2}\sum_{l=1}^{n}\hat{g}_{jl}(\frac{1}{k})$$

where C_i's and A are computed for some value of j.

In case $n=2$, ($Y=\{y\}$, $\hat{F}_1(Y)=\hat{F}'(y)$) we have only one equation

$$-y\hat{F}'(y)+\hat{F}'(y)=\frac{1}{y}C_1+A \tag{A.30}$$

Thus $F(y)=Cy^{C_1}(1-y)^{-(A+C_1)}$ (i.e. F is of Dirichlet form).

In case $n\ge 3$, by subtracting two Equations A.28 for $1\le i_1,i_2\le n-1$, $i_1\ne i_2$, we get:

$$\hat{F}_{i_1}(Y)-\hat{F}_{i_2}(Y)=\frac{C_{i_1}}{y_{i_1}}-\frac{C_{i_2}}{y_{i_2}} \tag{A.31}$$

The general solution for Equations A.31 is:

$$\hat{F}(Y)=h(\sum_{i=1}^{n-1}y_i)+\sum_{i=1}^{n-1}C_i\ln y_i \tag{A.32}$$

Substituting the solution into Equation A.28, and solving for h, we get:

$$F(Y) = C\prod_{i=1}^{n} y_i^{C_i} \tag{A.33}$$

where C is an arbitrary constant, $y_n = 1 - \sum_{i=1}^{n-1} y_i$ and $C_n = -A - \sum_{i=1}^{n-1} C_i$.

Similarly, $G(X) = B\prod_{j=1}^{k} x_j^{B_j}$, and by substitution $y_i = \frac{1}{n}$ ($x_j = \frac{1}{n} z_{j\cdot}$, $w_{ji} = \frac{z_{ji}}{z_{j\cdot}}$ where $z_{j\cdot} = \sum_{i=1}^{n} z_{ji}$) Equation A.26 transforms to:

$$\begin{array}{l} \sum_{l=1}^{n-1}\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)(C_l - C_n) - \left(\frac{B_j}{z_{j\cdot}} - \frac{B_k}{z_{k\cdot}}\right) = \\ \qquad \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] - \sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\hat{g}_{jl}(Z) \end{array} \tag{A.34}$$

By substitution we can see that $\hat{g}_p(Z) = \sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ji}\ln z_{ji}$ is a particular solution of Equation A.34. From Equations A.29 and A.34, the coefficients $\{\alpha_{ji}\}$, $\{C_i\}$ and $\{B_j\}$ satisfy: $C_i = \sum_{m=1}^{k}\alpha_{mi}$, $B_j = \sum_{l=1}^{n}\alpha_{jl}$, $B_k = \frac{1}{k}\sum_{m=1}^{k} B_m$ and, by symmetry, $C_n = \frac{1}{n}\sum_{l=1}^{n} C_l$ (see Section A.2.1 for derivations)

We must now solve the homogeneous first-order partial differential equation:

$$\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] - \sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\hat{g}_{jl}(Z) = 0 \tag{A.35}$$

Multiplying each equation by $z_{j\cdot}$, taking the sum by j, $1 \le j \le k-1$ and simplifying we get:

$$\sum_{l=1}^{n} z_{kl}\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = 0 \tag{A.36}$$

so Equations A.35 transforms to (for $1 \le j \le k-1$):

$$\sum_{l=1}^{n} z_{jl}\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = \sum_{l=1}^{n} z_{jl}\hat{g}_{jl}(Z) \tag{A.37}$$

Substitution $\hat{g}(Z) = \bar{g}(\bar{Z})$, where $\bar{z}_{ji} = \frac{z_{ji}}{z_{j+1,i}}$ for $1 \le j \le k-1$, $1 \le i \le n$ gives:

$$z_{ji}\hat{g}_{ji}(Z) = \begin{cases} \bar{z}_{ji}\bar{g}_{ji}(\bar{Z}) + \frac{z_{ji}}{z_{ki}}\bar{z}_{k-1,i}\bar{g}_{k-1,i}(\bar{Z}) & j = 1 \\ -\bar{z}_{j-1,i}\bar{g}_{j-1,i}(\bar{Z}) + \bar{z}_{ji}\bar{g}_{ji}(\bar{Z}) + \frac{z_{ji}}{z_{ki}}\bar{z}_{k-1,i}\bar{g}_{k-1,i}(\bar{Z}) & 2 \le j \le k-1 \end{cases} \tag{A.38}$$

and

$$\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = \frac{1}{z_{kl}}\bar{z}_{k-1,l}\bar{g}_{k-1,l}(\bar{Z}) \tag{A.39}$$

Plugging Equations A.38 and A.39 into Equations A.37 gives

$$\begin{array}{ll} \sum_{l=1}^{n}\bar{z}_{1l}\bar{g}_{1l}(\bar{Z}) = 0 & j = 1 \\ \sum_{l=1}^{n}\bar{z}_{jl}\bar{g}_{jl}(\bar{Z}) - \sum_{l=1}^{n}\bar{z}_{j-1,l}\bar{g}_{j-1,l}(\bar{Z}) = 0 & 2 \le j \le k-1 \end{array} \tag{A.40}$$

Thus for all j, $(1 \leq j \leq k-1)$:

$$\sum_{l=1}^{n} \bar{z}_{jl}\bar{g}_{jl}(\bar{Z}) = 0 \tag{A.41}$$

Additional substitution $\bar{g}(\bar{Z}) = \tilde{g}(\tilde{Z})$, where

$$\begin{array}{ll} \tilde{z}_{ji} = \frac{\bar{z}_{ji}}{\bar{z}_{j,i+1}} & 1 \leq i \leq n-1,\ 1 \leq j \leq k-1 \\ \tilde{z}_{jn} = \bar{z}_{jn} & 1 \leq j \leq k-1 \end{array} \tag{A.42}$$

gives (for $1 \leq j \leq k-1$):

$$\tilde{z}_{jn}\tilde{g}_{jn}(\tilde{Z}) = 0 \tag{A.43}$$

The general solution to Equations A.43 is

$$\tilde{g}(\tilde{Z}) = h(\{\tilde{z}_{ji} | 1 \leq i \leq n-1,\ 1 \leq j \leq k-1\}) + C_h \tag{A.44}$$

where h is an arbitrary differentiable function of $(n-1) \times (k-1)$ variables and C_h is an arbitrary constant.

Thus the general solution for g of the functional equation 4.5 is:

$$g(Z) = C \left[\prod_{i=1}^{n} \prod_{j=1}^{k} z_{ji}^{\alpha_{ji}} \right] H \left(\left\{ \frac{z_{ji} z_{j+1,i+1}}{z_{j+1,i} z_{j,i+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1 \right\} \right) \tag{A.45}$$

where H is an arbitrary differentiable function of $(n-1) \times (k-1)$ variables and C is an arbitrary constant.

Using $z_{ji} = \theta_{j|i} = \frac{\theta_{ij}}{\theta_{i\cdot}}$, $y_i = \theta_{i\cdot}$ (for $1 \leq i \leq n$, $1 \leq j \leq k$), as well as properties of the particular solution of Equation A.34 ($C_i = \sum_{m=1}^{k} \alpha_{mi}$) we get from Equations A.33,A.45:

$$p(\{\theta_{ij}\}) = C \left[\prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{\alpha_{ij}} \right] H \left(\left\{ \frac{\theta_{ij} \theta_{i+1,j+1}}{\theta_{i+1,i} \theta_{i,j+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1 \right\} \right) \tag{A.46}$$

This solves Equation 4.1 and completes the proof of Theorem 4. ■

A.2.1 Some technical details

In this appendix we want to show that $\hat{g}_p(Z) = \sum_{i=1}^{n} \sum_{j=1}^{k} \alpha_{ji} \ln z_{ji}$ is a particular solution of Eq A.34, where the coefficients $\{\alpha_{ji}\}$ should satisfy some constrains imposed by Eq A.34 and Eq A.29. Note that $\frac{\partial \hat{g}_p}{\partial z_{ji}}(Z) = \frac{\alpha_{ji}}{z_{ji}} - \frac{\alpha_{ki}}{z_{ki}}$.

Substituting $\hat{g}_p(Z)$ into right hand side of Eq A.34, we get:

$$\begin{aligned}
&\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\left(\frac{\alpha_{ml}}{z_{ml}}-\frac{\alpha_{kl}}{z_{kl}}\right)\right]-\sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\left(\frac{\alpha_{jl}}{z_{jl}}-\frac{\alpha_{kl}}{z_{kl}}\right)\\
&=\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)\left(\sum_{m=1}^{k-1}\alpha_{ml}-(1-z_{kl})\frac{\alpha_{kl}}{z_{kl}}\right)\right]-\sum_{l=1}^{n}\left(\frac{\alpha_{jl}}{z_{j\cdot}}-\frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}}\right)\\
&=\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right]-\sum_{l=1}^{n}\frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}}+\sum_{l=1}^{n}\frac{\alpha_{kl}}{z_{k\cdot}}-\sum_{l=1}^{n}\frac{\alpha_{jl}}{z_{j\cdot}}+\sum_{l=1}^{n}\frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}}\\
&=\sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right]-\left((1-\frac{z_{jn}}{z_{j\cdot}})-(1-\frac{z_{kn}}{z_{k\cdot}})\right)\sum_{m=1}^{k}\alpha_{mn}\\
&\quad-\left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}}-\frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)\\
&=\sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right]-\left(\frac{\sum_{l=1}^{n-1}z_{jl}}{z_{j\cdot}}-\frac{\sum_{l=1}^{n-1}z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{mn}\\
&\quad-\left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}}-\frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)\\
&=\sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}}-\frac{z_{kl}}{z_{k\cdot}}\right)(\sum_{m=1}^{k}\alpha_{ml}-\sum_{m=1}^{k}\alpha_{mn})\right]-\left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}}-\frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)
\end{aligned}\tag{A.47}$$

Thus, from Eq A.47 and Eq A.34, we get:

$$\begin{array}{ll}
C_i - C_n = \sum_{m=1}^{k}\alpha_{mi}-\sum_{m=1}^{k}\alpha_{mn} & 1\le i\le n-1\\
B_j = \sum_{l=1}^{n}\alpha_{jl} & 1\le j\le k
\end{array}\tag{A.48}$$

If $\hat{g}_p(Z)$ is indeed a particular solution to Eq A.34 then $\{\alpha_{ji}\}$ and $\{C_i\}$ should satisfy Eq A.29. Substituting $\hat{g}_p(Z)$ into Eq A.29 and using Eq A.48, we get $(1\le i\le n-1)$:

$$\begin{aligned}
C_i &= \sum_{m=1}^{k-1}(\alpha_{mi}-\alpha_{ki})+\frac{k}{2}(\alpha_{ji}+\alpha_{ki})-\frac{k}{2}(\alpha_{ji}-\alpha_{ki})=\sum_{m=1}^{k}\alpha_{mi}\\
A &= -k\sum_{l=1}^{n}\sum_{m=1}^{k-1}(\alpha_{ml}-\alpha_{kl})-\frac{k}{2}(B_j+B_k)+\frac{k}{2}\sum_{l=1}^{n}(\alpha_{jl}-\alpha_{kl})\\
&= -k\sum_{m=1}^{k}(B_m-B_k)-kB_k
\end{aligned}\tag{A.49}$$

Thus, $C_n=\sum_{m=1}^{k}\alpha_{mn}$ (from Eqs A.48,A.49) and using the fact that $A = -\sum_{l=1}^{n}C_l$ (from the definition of C_n, after Eq A.33), we get:

$$k\sum_{m=1}^{k}(B_m-B_k)+kB_k=\sum_{l=1}^{n}C_l \tag{A.50}$$

Thus $B_k=\frac{1}{k}\sum_{m=1}^{k}B_m$ and, by symmetry, $C_n=\frac{1}{n}\sum_{l=1}^{n}C_l$.

A.3 The Proof of Theorem 7

Lemma 5 states that in order to find all distributions that are dictated by global independence assumption it is enough to find a general solution for Equation 4.7. Equation 4.7 satisfies the scale-independence conditions of Lemma 10, thus its general solution can be found by treating $\{\theta_{\vec{x}}\}$ as independent variables.

Equations 4.7 for the binary $\mathbf{X}$ can be written as (normalization constant is inside H_i):

$$p(\{\theta_{\vec{x}}\}_{\vec{x}\in\{0,1\}^n})=\prod_{\vec{x}\in\{0,1\}^n}\theta_{\vec{x}}^{\alpha_{\vec{x},i}}H_i\left(\left\{\frac{\theta_{[\vec{x}\backslash x_i]=\vec{y},x_i=0}\theta_{[\vec{x}\backslash x_i]=\vec{y}+1,x_i=1}}{\theta_{[\vec{x}\backslash x_i]=\vec{y},x_i=1}\theta_{[\vec{x}\backslash x_i]=\vec{y}+1,x_i=0}}\right\}\right)\tag{A.51}$$

Applying logarithm to both sides of Equations 4.7 and to arguments of H_i and changing the variables to $\{\ln \theta_{\vec{x}}\}_{\vec{x}\in\{0,1\}^n}$ we get Equations 4.8 with:

$$\vec{b}_{ij}[m] = \begin{cases} 1 & [\vec{x}\setminus x_i] = \vec{y}, x_i = 0 \text{ or } [\vec{x}\setminus x_i] = \vec{y}+1, x_i = 1 \\ -1 & [\vec{x}\setminus x_i] = \vec{y}, x_i = 1 \text{ or } [\vec{x}\setminus x_i] = \vec{y}+1, x_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.52}$$

where $1 \le i \le n$, $0 \le j \le 2^{n-1}-2$, $0 \le m \le 2^n - 1$ and $\vec{y}, \vec{x}$ denote the binary representations of j and m respectively. The application of Lemma 6 and the following lemma (for $k = n$) concludes the proof. ■

Lemma 11 *Given $\vec{b}_{ij}$ as specified by Equation A.52,*

$$\bigcap_{i=1}^{k} B_i = A_k, \quad \text{for } k = 2, \ldots, n \tag{A.53}$$

where $B_i = span(\vec{b}_{i,0}, \ldots, \vec{b}_{i,2^{n-1}-2})$ and $A_k = span(\vec{a}_{k,0}, \ldots, \vec{a}_{k,2^{n-i}-1})$, s.t.

$$\vec{a}_{k,\vec{y}}[\vec{x}] = \begin{cases} 1 & \vec{x}_{\perp[k+1,\ldots,n]} = \vec{y}, \vec{x}^2_{\perp[1,\ldots,k]}\%2 = 0 \\ -1 & \vec{x}_{\perp[k+1,\ldots,n]} = \vec{y}, \vec{x}^2_{\perp[1,\ldots,k]}\%2 = 1 \\ 0 & \textit{otherwise.} \end{cases} \tag{A.54}$$

We use $\vec{y}$ and $\vec{x}$ to denote binary representation of indexes: $0 \le \vec{y} \le 2^{n-k}-1$ and $0 \le \vec{x} \le 2^n - 1$.

Proof: By induction on k.

Induction Basis: Consider $\mathcal{M}_1$,$\mathcal{M}_2$ matrices with rows $\{\vec{b}_{1,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{b}_{2,j}\}_{j=0}^{2^{n-1}-2}$ respectively:

$$\mathcal{M}_1 = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & & \end{bmatrix} \tag{A.55}$$

and

$$\mathcal{M}_1 = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & & \end{bmatrix}. \tag{A.56}$$

Note that the same pattern as in the first two rows, repeats itself with 4-column shift for all $2^{n-1}-1$ rows of $\mathcal{M}_1$ and $\mathcal{M}_2$ except the very last row. Considering the matrix $\mathcal{M} = [\mathcal{M}_1^T|\mathcal{M}_2^T]^T$, with row space equal to B_1+B_2, we can see that it is exactly of rank $3 \cdot 2^{n-2}-2$, since rows $\{\vec{b}_{1,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{b}_{2,j}\}_{j=1,3,\ldots,2^{n-1}-3}$ are linearly independent and $\vec{b}_{1,j} = \vec{b}_{2,j}$ for $j = 0, 2, \ldots, 2^{n-1}-2$. Thus, $dim(B_1 \cap B_2) = dim(B_1) + dim(B_2) - dim(B_1 + B_2) = (2^{n-1}-1) + (2^{n-1}-1) - (3 \cdot 2^{n-2} - 2) = 2^{n-2}$ and the basis for $B_1 \cap B_2$ is specified by $\vec{a}_{2,i} = \vec{b}_{1,2i}$, for $i = 0, \ldots, 2^{n-2}-1$.

Induction Step: We will demonstrate the correctness of the induction step for $k = 3$. The proof is easily extended for arbitrary k.

Consider matrices $\mathcal{M}_A$,$\mathcal{M}_3$ with rows $\{\vec{a}_{2,j}\}_{j=0}^{2^{n-2}-1}$ and $\{\vec{b}_{3,j}\}_{j=0}^{2^{n-1}-2}$ respectively, i.e.

$$\mathcal{M}_A = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & & & & & & \end{bmatrix} \tag{A.57}$$

and $\mathcal{M}_3$ is

$$\begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & & & & & & & & \end{bmatrix}. \tag{A.58}$$

The first 4-row pattern of $\mathcal{M}_3$ repeats itself with 8 column shift for all $2^{n-1}-1$ rows of $\mathcal{M}_3$ (the last pattern is truncated to 3 rows). For a general $\mathcal{M}_k$ matrix - the pattern of the first 2^{k-1} rows will repeat itself with shift of 2^k columns, which is twice the length of non-zero segment in $\{\vec{a}_{k-1,i}\}$.

It is clear now, that for the matrix $\mathcal{M} = [\mathcal{M}_A^T | \mathcal{M}_3^T]^T$ the rows $\{\vec{b}_{3,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{a}_{2,j}\}_{j=1,3,\ldots,2^{n-2}-1}$ are linearly independent and $\vec{a}_{2,j} = \vec{b}_{3,4j} - \vec{b}_{3,4j+2} + \vec{a}_{2,j+1}$, for $j = 0, 2, \ldots, 2^{n-2}-2$. Thus, $dim(A_2 \cap B_3)$ is equal to number of dependent rows in $\mathcal{M}$, i.e. 2^{n-3}, and the basis for A_3 is $\vec{a}_{3,j} = \vec{a}_{2,2j} - \vec{a}_{2,2j+1} = \vec{b}_{3,2j} - \vec{b}_{3,2j+2}$, for $j = 0, \ldots, 2^{n-3}-1$.

For the general $k = 3, \ldots, n$, the rows $\{\vec{b}_{k,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{a}_{k,j}\}_{j=1,3,\ldots}^{2^{n-(k-1)}-1}$ are linearly independent and $\vec{a}_{k,j} = \vec{a}_{k-1,2j} - \vec{a}_{k-1,2j+1} = \sum_{l=0}^{2^{k-2}-1} (-1)^l \vec{b}_{k,2j+2l}$ for $j = 0, \ldots, 2^{n-k}-1$. ∎