The Annals of Statistics
1997, Vol. 25, No. 3, 1344–1369

A CHARACTERIZATION OF THE DIRICHLET DISTRIBUTION THROUGH GLOBAL AND LOCAL PARAMETER INDEPENDENCE[1]

BY DAN GEIGER[2] AND DAVID HECKERMAN

Technion and Microsoft Research

We provide a new characterization of the Dirichlet distribution. Let $\theta_{ij}, 1 \le i \le k, 1 \le j \le n$, be positive random variables that sum to unity. Define $\theta_{i\cdot} = \sum_{j=1}^{n} \theta_{ij}$, $\theta_{I\cdot} = \{\theta_{i\cdot}\}_{i=1}^{k-1}$, $\theta_{j|i} = \theta_{ij}/\sum_j \theta_{ij}$ and $\theta_{J|i} = \{\theta_{j|i}\}_{j=1}^{n-1}$. We prove that if $\{\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|n}\}$ are mutually independent (where $\theta_{\cdot J}$ and $\theta_{I|j}$ are defined analogously), and each parameter set has a strictly positive pdf, then the pdf of θ_{ij} is Dirichlet. This characterization implies that under assumptions made by several previous authors for selecting a Bayesian network structure out of a set of candidate structures, a Dirichlet prior on the parameters is inevitable.

1. Introduction. A statistical model that represents a large collection of discrete random variables imposes severe computational complexity unless some notion of independence is introduced that decreases the dimensionality of the model. Graphical models address this problem. A graphical model represents a collection of random variables by a graph; each node in the graph represents a random variable, and the lack of an edge between two nodes represents a conditional independence assertion. Such models have been extensively studied in the fields of statistics (e.g., [17, 34, 18, 15, 31, 6, 25]), artificial intelligence and computer science (e.g., [20, 21, 10, 23]), operations research (e.g., [28, 29]) and philosophy (e.g., [32]). For an introduction to graphical models, see [22, 35] and references therein.

Graphical models are based on directed acyclic graphs, undirected graphs or a combination thereof. A class of graphical models that is based on directed acyclic graphs, called *Bayesian networks*, is the most suitable among current graphical models to be constructed from expert knowledge rather than from sampling data. Each node i in a Bayesian network represents a random variable s_i and the joint distribution satisfies

$$p(s_1, \dots, s_n) = \prod_i p\left(s_i | s_{i_1}, \dots s_{i_k}\right),$$

where $i_1, \dots, i_k$ are nodes from which a directed edge is drawn into node i. These nodes are called the *parents* of i. A simple example of a Bayesian

Received February 1995; revised April 1996.

[1] A preliminary version of this article appears in [9].

[2] Part of this work was done while the author visited Microsoft Research Center.

AMS 1991 *subject classifications*. Primary 62E10, 60E05; secondary 62A15, 62C10, 39B99.

Key words and phrases. Bayesian network, characterization, Dirichlet distribution, functional equation, graphical model, hyper-Markov law.

network is the well-known Markov chain over $\{s_i | 1 \leq i \leq n\}$, which represents the distribution $p(s_1, \ldots, s_n) = \prod_i p(s_i | s_{i-1})$. Because the joint distribution is composed of local conditional probability tables between closely related variables, these tables can often be assessed directly from experts. Consequently Bayesian networks have become the dominant model in artificial intelligence for representing knowledge needed for reasoning tasks that require the explicit representation of uncertainty.

In recent years, researchers have realized that, although Bayesian networks can be constructed directly from expert knowledge (often using advanced computerized elicitation techniques [11]), it is advantageous to use data to update both the parameters and structure of a graphical model. The latter problem has been addressed by several researchers who have investigated Bayesian methods for model averaging and selection when the models are Bayesian networks [2, 4, 12, 30]. Such a task is often referred to as *learning*. These approaches all have the same basic components: a scoring rule and a search procedure. The scoring rule takes data and a network structure and returns a score reflecting the goodness-of-fit of the data to the structure. A search procedure generates networks for evaluation by the scoring rule. These approaches use the two components to identify a network structure or set of structures that can be used, for example, to predict future observations.

Suppose we have a set of discrete random variables $\{s_1, \ldots, s_n\} = U$, and a data set $D = \{C_1, \ldots, C_m\}$ where each case (i) is an instance of some or of all the variables in U. Let B be a Bayesian network structure (a directed acyclic graph) and B^h stand for the hypothesis corresponding to B (see Section 3). An important quantity for both model averaging and model selection is the posterior probability of B^h given D, $p(B^h | D) = cp(B^h)p(D | B^h)$, where c is a normalizing factor.

To compute $p(D | B^h)$ in closed form, researchers have made several assumptions. One, the prior probability of each structure is positive—that is, $p(B^h) > 0$ for every B. Two, for each network structure, the parameters associated with each node are mutually independent (global parameter independence [31]), and the parameters associated with a node and each instance of its parents are mutually independent (local parameter independence [31]). Three, if a node has the same parents in two distinct networks structures, then the prior distribution of the parameters associated with this node are identical for both structures (parameter modularity [12]). Four, each case is complete—namely, each case is an instance of all the variables represented by the network. Five, the prior distribution of the parameters associated with each node and each instance of its parents is Dirichlet. The last two assumptions are made so as to create a conjugate sampling situation. Namely, after data is seen, the distributions of the parameters stay in the same family—the Dirichlet family.

The contribution of this article is a characterization of the Dirichlet distribution based on local and global parameter independence, and on the assumption that the prior distributions of all the parameters are strictly positive pdfs. In Section 3, we explore the circumstances under which our characterization implies that the distribution of the parameters associated with each node in a Bayesian network must be Dirichlet, in which case the fifth assumption for learning is redundant. The assumption of parameter modularity, which is further discussed in Section 3, plays a key role in learning Bayesian networks, but is not needed for the characterization theorem. Consequently, the characterization can be described more easily without reference to graphical models as follows.

Suppose s and t are two discrete random variables having finite domains, $\{s_i\}_{i=1}^{k}$ and $\{t_j\}_{j=1}^{n}$, respectively. We wish to infer an unrestricted joint probability $p(s,t)$ from a sample of pairs of values (s_i, t_j) of s and t. A Bayesian approach to this statistical inference problem is to associate with $p(s_i, t_j)$ a (multinomial) parameter θ_{ij}, assign $\{\theta_{ij} | 1 \leq i \leq k, 1 \leq j \leq n\}$ a prior joint pdf and compute the posterior joint pdf of $\{\theta_{ij}\}$ given the observed set of pairs of values. There are two alternatives to this approach that can be described as follows.

Let $\theta_{i\cdot} = \sum_{j=1}^{n} \theta_{ij}$ stand for the parameter associated with $p(s = s_i)$, and let $\theta_{j|i} = \theta_{ij} / \sum_j \theta_{ij}$ stand for the parameter asociated with $p(t = t_j | s = s_i)$. Furthermore, let $\theta_{I\cdot} = \{\theta_{i\cdot}\}_{i=1}^{k-1}$ and $\theta_{J|i} = \{\theta_{j|i}\}_{j=1}^{n-1}$. We assume that $\{\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}\}$ are mutually independent and that each has a prior pdf. According to Bayesian practice, we compute the joint posterior appropriately —that is, we update the pdf for $\theta_{I\cdot}$ according to the counts of $s = s_i$ in the observed pairs, and update the pdf of $\theta_{J|i}$ according to the counts of $t = t_j$ in all pairs in which $s = s_i$. In a symmetric fashion, let $\theta_{\cdot j} = \sum_{i=1}^{k} \theta_{ij}$, $\theta_{i|j} / \sum_i \theta_{ij}$, $\theta_{\cdot J} = \{\theta_{\cdot j}\}_{j=1}^{n-1}$ and $\theta_{I|j} = \{\theta_{i|j}\}_{i=1}^{k-1}$. We assume that $\{\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|n}\}$ are mutually independent, and that each set of parameters has a prior pdf. We compute the posterior pdf for $\theta_{\cdot J}$ according to the counts of $t = t_j$, and the posterior pdf of $\theta_{I|j}$ according to the counts of $s = s_i$ in all pairs in which $t = t_j$.

To make these techniques operational, one must choose a specific prior pdf for the multinomial parameters. The standard choice of a pdf for $\{\theta_{ij}\}$—typically made for practical reasons—is a Dirichlet distribution. When such a choice is made, it can be shown that $\{\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}\}$ are mutually independent and each parameter set has a prior Dirichlet pdf, and (similarly) that $\{\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|n}\}$ are mutually independent and each parameter set has a prior Dirichlet pdf.

The result proved in this article is that under these independence assumptions and the assumption that each parameter set has a strictly positive pdf, a prior Dirichlet pdf for $\{\theta_{ij}\}$ is the only possible choice. We conjecture that the assumption of strict positivity can be dropped without affecting the conclusion. In Section 2, we discuss our proof technique, which uses the tool of functional equations. We also review briefly the applicability of this technique to other characterization problems in statistics. In Section 3, we discuss an extension of our characterization from two-way tables to n-way tables, as well as the implications of our characterization for learning Bayesian networks. Further extensions are described in Section 4. An analogous result that characterizes the normal-Wishart distribution is outlined in [9].

2. Background and technical summary. The Dirichlet pdf is defined as follows. Let $\phi_1, \dots, \phi_l$ be positive random variables that sum to 1. Then $\phi_1, \dots, \phi_{l-1}$ have a Dirichlet pdf f if

$$f(\phi_1, \dots, \phi_{l-1}) = \frac{\Gamma\left(\sum_{i=1}^{l} \alpha_i\right)}{\prod_{i=1}^{l} \Gamma(\alpha_i)} \prod_{i=1}^{l} \phi_i^{\alpha_i - 1}, \tag{1}$$

where $\phi_l = 1 - \sum_{i=1}^{l-1} \phi_i$ and α_i are positive hyperparameters (See, e.g., [7], [36]).

We use the following conventions. Suppose $\{\theta_{ij}\}$, $1 \le i \le k, 1 \le j \le n$, is a set of positive random variables that sum to 1. Let $\theta_{i\cdot}$, $\theta_{\cdot J}$, $\theta_{I\cdot}$, $\theta_{\cdot J}$, $\theta_{j|i}$, $\theta_{i|j}$, $\theta_{J|i}$ and $\theta_{I|j}$ be defined as in the introduction. Consequently, $\theta_{i\cdot}\theta_{j|i} = \theta_{\cdot j}\theta_{i|j}$ for every i and j. Let f_U be the joint pdf of $\{\theta_{ij}\}$, f_I be the pdf of $\theta_{I\cdot}$, and $f_{J|i}$ be the pdf of $\theta_{J|i}$. Similarly, let f_J be the pdf of $\theta_{\cdot J}$, and $f_{I|j}$ be the pdf of $\theta_{I|j}$. Finally, let f_{IJ} be the joint pdf of $\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}$ and f_{JI} be the joint pdf of $\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|n}$.

A Dirichlet pdf for $\{\theta_{ij}\}$ is given by

$$f_U(\{\theta_{ij}\}) = c \prod_{i=1}^{k} \prod_{j=1}^{n} \theta_{ij}^{\alpha_{ij} - 1}, \tag{2}$$

where $\theta_{kn} = 1 - \sum_A \theta_{ij}$, $A = \{(i, j) | 1 \le i, j \le n, i \ne k \text{ or } j \ne n\}$, c is the normalization constant and α_{ij} are positive constants.

We observe that f_U and f_{IJ} are related through a change of variables. Because both $\{\theta_{i\cdot}\}_{i=1}^{k}$ and $\{\theta_{j|i}\}_{j=1}^{n}$ are defined in terms of $\{\theta_{ij}\}$, and because $\theta_{ij} = \theta_{i\cdot}\theta_{j|i}$, there exists a one-to-one and onto correspondence between $\{\theta_{ij}\}$ and $\{\theta_{i\cdot}\} \cup \{\theta_{j|i}\}$. The Jacobian $J_{k,n}$ of this transformation is given by

$$J_{kn} = \prod_{i=1}^{k} \theta_{i\cdot}^{n-1} \tag{3}$$

(see [12]).

The following lemma provides a known property of the Dirichlet distribution. A slightly different version is stated in [6], Lemma 7.2.

LEMMA 1. *Let $\{\theta_{ij}\}, 1 \le i \le k, 1 \le j \le n$, where k and n are integers greater than* 1, *be a set of positive random variables having a Dirichlet distribution. Then, $f_I(\theta_{I\cdot})$ is Dirichlet, $f_{J|i}(\theta_{J|i})$ is Dirichlet for every $i, 1 \le i \le k$, and $\{\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}\}$ are mutually independent.*

PROOF. Set $\theta_{ij} = \theta_{i\cdot}\theta_{j|i}$ in (2), multiply by J_{kn}, and regroup terms. □

The main claim of this article is that, under the assumption of a strictly positive pdf for $\{\theta_{ij}\}$, the converse holds as well. More specifically, we prove the following theorem (the proof is given in the Appendix).

THEOREM 2. *Let* $\{\theta_{ij}\}, 1 \le i \le k, 1 \le j \le n, \sum_{ij} \theta_{ij} = 1$, *where* k *and* n *are integers greater than* 1, *be positive random variables having a strictly positive pdf* $f_U(\{\theta_{ij}\})$. *If* $\{\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}\}$ *are mutually independent and* $\{\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|n}\}$ *are mutually independent, then* $f_U(\{\theta_{ij}\})$ *is Dirichlet.*

Recall that f_U can be written both in terms of f_{IJ} and in terms of f_{JI} by a change of variables and using the Jacobian given by (3). Because both representations must be equal, and using the independence assumptions made by Theorem 2 to factor f_{IJ} and f_{JI}, we get the equality,

$$(4)\quad \left(\prod_{j=1}^{n} \theta_{\cdot j}^{k-1}\right)^{-1} f_J(\theta_{\cdot J}) \prod_{j=1}^{n} f_{I|j}(\theta_{I|j}) = \left(\prod_{i=1}^{k} \theta_{i\cdot}^{n-1}\right)^{-1} f_I(\theta_{I\cdot}) \prod_{i=1}^{k} f_{J|i}(\theta_{J|i}).$$

This equality, which is a functional equation, summarizes the independence assumptions stated in Theorem 2.

Methods for solving functional equations such as 4, that is, finding all functions that satisfy them under different regularity assumptions, are discussed in [1]. We use the following technique. First, we argue that any positive solution to (4) must be differentiable in any order ([1], Section 4.2.2, "Deduction of differentiability from integrability"). Then we take repeated derivatives of (4) and obtain a differential equation, the solution of which after appropriate specialization is the general solution of (4) ([1], Section 4.2, "Reduction to differential equations").

For example, to demonstrate that the only differentiable functions that satisfy $f(x + y) = f(x) + f(y)$ are linear, one can take a derivative wrt (with respect to) x and obtain $f'(x + y) = f'(x)$. Because the latter equality holds for all y, it follows that $f'(t)$ is constant and thus $f(t)$ is linear in t [1]. This functional equation is one of Cauchy's fundamental equations and it establishes the memoryless property that characterizes the exponential distribution (e.g., [19]). In (4), there are several functions and several free variables, the number of which depends on n and k. For example, when $n = k = 2$ and by renaming of variable and function names, (4) can be written as follows:

$$(5)\qquad f_0(y) g_1(z) g_2(w) = g_0(x) f_1\left(\frac{yz}{x}\right) f_2\left(\frac{y(1-z)}{1-x}\right),$$

where

$$x = yz + (1-y)w$$

and where y, z and w replace $\theta_{\cdot j=1}, \theta_{i=1|j=1}, \theta_{i=1|j=2}$, respectively. The solution of this equation is given in the Appendix.

Járai [13] has extensively investigated the following type of functional equations:

$$(6)\qquad f(t) = h(t, y, f_0(y), f_1(g_1(t, y)), \dots, f_n(g_n(t, y))),$$

where $f, f_0, \dots, f_n$ are unknown functions, $h, g_1, \dots, g_n$ are known functions satisfying some regularity conditions and all variable and function values may be vectors. Our functional equation, as well as many other functional equations, can be written in this form. Járai showed that every measurable solution of this equation must be continuous (Theorem 3.3 in [13]). Because (4) can be written in this form and does satisfy the needed regularity conditions, we may conclude that any pdf that solves it must be continuous (because a pdf is Lebesgue integrable and thus measurable).

Járai has also dealt with functional equations of the type

$$f(t) = \sum_{i=1}^{n} h_i(t, y, f_i(g_i(t, y))), \tag{7}$$

where f and f_i are the unknown functions. Note that this equation is a special case of (6). For this type of equation, Járai proved, under regularity conditions on the known functions g_i and h_i (which hold in our case), that any continuous solution must be indefinitely differentiable (Theorems 5.2, 7.2 in [13]). Actually some stronger results of this sort are proved in [13]. Thus, for example, the above theorems imply that any measurable solution of $f(x + y) = f(x) + f(y)$ must have a first derivative and so we are allowed to take a derivative of this equation; therefore, all measurable solutions are linear.

In solving (4), we can use the first part of Járai's contribution and obtain continuity. To apply the second part, we take the logarithm of the equation and obtain a functional equation of the form of (7). We assume that the solutions are strictly positive and measurable. Because the logarithm of a positive measurable function is measurable, we can now use Járai's theorems and obtain that all positive measurable solutions of (4) have infinitely many derivatives.

Járai's theorems are very useful in statistical applications because they "upgrade" results proved for smooth pdfs to any pdfs. We shall now demonstrate their usefulness for another well-known characterization of the Dirichlet distribution due to Darroch and Ratcliff [5]. Their bivariate theorem states:

> *Let X and Y be two continuous, positive random variables which satisfy the inequality $X + Y < s$. Assume the pdf of X and Y on $(0, s)$ and $X/s - Y$ and $Y/s - X$ on $(0, 1)$ are all continuous. Then, if and only if $X/s - Y, Y$ are independent and $Y/s - X, X$ are independent, X, Y have a Dirichlet pdf.*

This theorem is similar in flavor to Theorem 2 because it also merely assumes independence assumptions on some transformation of the given random variables. The difference is the transformation. In Theorem 2, the transformations arise from the use of a Dirichlet pdf as a prior distribution of multinomial parameters while the Darroch and Ratcliff bivariate theorem is derived from conditions of neutrality. Nevertheless, Járai's theorems are applicable also for the latter problem. As Darroch and Ratcliff do, the joint pdf of X and Y can be written in two distinct ways. Equating these representations, as in (4), forms a functional equation. This functional equation is of the type dealt with by Járai and consequently, the theorem by Darroch and

Ratcliff holds even without assuming continuous pdfs. Indeed, among other results, this was shown, using other techniques, by [8, 14]. Note that if a pdf is in fact a gpdf, that is, it contains a discrete element, then Lebesgue integrability is not satisfied and this technique is not applicable as is. In this case, one may resort to the functional equations defined by the characteristic functions near the origin. A review of many characterization problems in statistics can be found in [16, 26, 27]. These texts do not use the elementary solution method used herein.

Another well-known characterization of the Dirichlet distribution which is described by several authors is based on W. E. Johnson's sufficientness postulate (See [37] and references therein). This characterization is based, loosely speaking, on the assumption that for exchangeable sequences the expectation of the parameter of the ith category depends only on the counts of the ith category and the total count. Our assumptions on the other hand, in particular, global parameter independence, were originally made so as to facilitate a prior-to-posterior analysis [6]. In this article, we show that these assumptions, as a by-product, also determine a restrictive class of prior pdfs. Clearly, any set of assumptions that yields a Dirichlet prior is doomed to be violated in a general setting because the class of Dirichlet priors is not expressive enough; for example, all members are unimodal and thus Dirichlet mixtures are sometimes preferable. This point is raised again in Section 3.

As a word of caution, one must realize that there are functional equations in statistics that include solutions which do not have a derivative. For example, in [19], the functional equation that defines a multivariate exponential distribution $F(X, Y)$ through an extended version of the memoryless property,

$$F(x_1 + y, x_2 + y) = F(x_1, x_2)F(y, y),$$

$x_1, x_2, y > 0$, yields, provided we assume that the marginals are exponential, a distribution function of the form,

$$P(X > x, Y > y) = \exp\{-\lambda_1 x - \lambda_2 y - \lambda_{12} \max\{x, y\}\},$$

which is not differentiable. By taking the logarithm and a derivative of this functional equation, we would have obtained the solution,

$$P(X > x, Y > y) = \exp\{-\lambda_1 x - \lambda_2 y\}$$

thus losing an important term of the general solution. This situation occurs because a regularity condition of Járai is violated [the rank of the matrix of the first derivatives wrt y,$(\partial g_i(t, y)/\partial y)$, for each g_i in (6) must equal the dimensionality of the domain of f_i; here, the mapping $y \to (y, y)$ fails to meet this condition because the rank is 1 rather than 2].

3. Implications for learning.

We now explain how our characterization applies to learning Bayesian networks. We concentrate on Bayesian networks for two discrete random variables s and t whose joint distribution is $p(s, t)$. There are three possible Bayesian network structures with two nodes: the structure that contains no edge between its two nodes s and t (B_0), the structure $s \to t$ (B_1) and the structure $t \to s$ (B_2). The structure B_0 corresponds to the assertion that s and t are independent, whereas the structures B_1 and B_2 correspond to the assertion that s and t are dependent; B_1 represents the factorization $p(s,t) = p(s)p(t|s)$, whereas B_2 represents the factorization $p(s,t) = p(t)p(s|t)$.

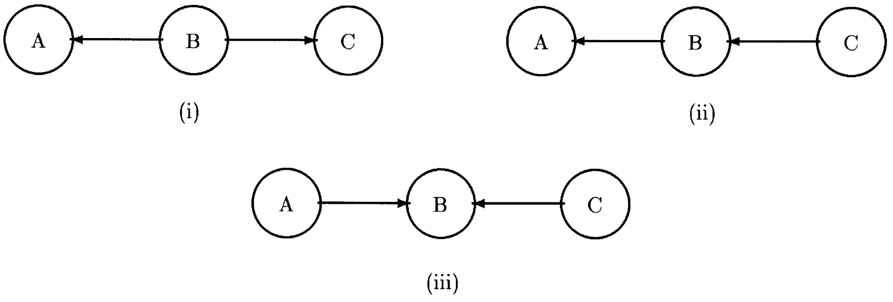
DEFINITION. Two Bayesian network structures B_1 and B_2 for a set of discrete random variables U are Markov *equivalent* if they encode the same set of independence assertions for U.

For example, network structures in which every pair of nodes are connected (complete network structures) are equivalent, because each such a network structure encodes no independence assertions for U. Another example is given in Figure 1. Characterizations of Markov equivalent Bayesian networks for discrete random variables are obtained in [3, 33].

Given a set of discrete random variables U having a joint pdf $p(U)$ and a network structure B, we define hypothesis B^h to be the hypothesis that precisely the independence assertions entailed by B hold in the joint distribution $p(U)$. By this definition of B^h, if network structures B_1 and B_2 are Markov equivalent, then $B_1^h = B_2^h$.

Recalling the notation introduced in Section 1, we have that $\theta_{i\cdot} = \sum_{j=1}^{n} \theta_{ij}$ denote the multinomial parameters associated with $p(s = s_i)$ and $\theta_{j|i} = \theta_{ij}/\sum_j \theta_{ij}$ denote the multinomial parameters associated with $p(t = t_j | s = s_i)$. Given $B_1^h = B_2^h$ and that s and t have a joint multinomial distribution, we obtain

$$f_{IJ}\left(\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k} | B_1^h\right) = f_{IJ}\left(\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k} | B_2^h\right),$$
$$f_{JI}\left(\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|k} | B_2^h\right) = f_{JI}\left(\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|k} | B_1^h\right)$$



(i)

(ii)

(iii)

FIG. 1. *The pair of network structures* (*i*) *and* (*ii*) *are equivalent because both encode precisely those distributions where A and C are conditionally independent, given B. Network structure* (*iii*) *is not equivalent to them, because it encodes those distributions where A and C are marginally independent.*

Using local and global parameter independence to factor f_{IJ} and f_{JI}, we immediately obtain (4). (We suppress the conditioning hypotheses because $B_1^h = B_2^h$.) Thus for the two complete network structures, the only possible strictly positive prior pdfs on their parameters is, according to Theorem 2, the Dirichlet distribution.

In this derivation, in order to apply Theorem 2, we assumed local and global parameter independence, a regularity condition that each B_i^h has a positive probability (because we condition on B_i^h), and that each f_{IJ} is a strictly positive pdf. Also, we used the equality $B_1^h = B_2^h$.

The parameter priors for the noncomplete network structure B_0 are determined from the added assumption of parameter modularity, which says that if the nodes corresponding to a random variable have the same parents in two different structures, then the prior pdfs associated with the parameter(s) of those nodes are equal. In our two-variable example, parameter modularity gives us

$$f_i\left(\theta_{i\cdot}|B_1^h\right) = f_i\left(\theta_{i\cdot}|B_0^h\right),$$

$$f_j\left(\theta_{\cdot j}|B_2^h\right) = f_j\left(\theta_{\cdot j}|B_0^h\right).$$

These equalities imply that the prior for each parameter set of B_0 is a Dirichlet distribution as well.

Recall that the hyperparameters of a Dirichlet distribution can be written as $N\alpha_{ij}$ where N is an *equivalent sample size* (the size of an imaginary set of complete cases that summarize a person's prior knowledge) and α_{ij} is the expectation of θ_{ij}. The equivalent sample size can be viewed as the assessor's confidence in the expectations of each θ_{ij}. A joint Dirichlet prior is therefore quite restrictive, because it accommodates only one equivalent sample size or confidence for the entire set of variables. Thus, a practical ramification of our characterization is that the commonly made global and local parameter independence assumption is inappropriate whenever a single equivalent sample size is not sufficient to describe prior knowledge. Such a situation occurs, for example, if knowledge about $\theta_{J|i}$ is more precise than knowledge about $\theta_{I\cdot}$.

The inevitable choice of a Dirichlet prior for two-variable networks (two-way tables) is easily generalized to the n-variate case by induction and without the need to solve additional functional equations. The inductive proof uses the fact that the sample space of a set of discrete random variables can be viewed as the sample space of a single discrete random variable. Here, we state the result in the notation of this article. For a proof, consult [12], Theorem 7.

Suppose $s_1, \dots, s_m$ are m discrete random variables having finite domains. With each of the n_j possible assignments of values to s_j, $j = 1, \dots, m$, we associate a multinomial parameter $\theta_{i_1,\dots,i_m}$. Analogously, to the case $m = 2$ discussed in previous sections, let

$$\theta_{i_1,\dots,i_j} = \sum_{i_{j+1},\dots,i_m} \theta_{i_1,\dots,i_m} \quad \text{and} \quad \theta_{i_j|i_1,\dots,i_{j-1}} = \theta_{i_1,\dots,i_j} \Big/ \sum_{i_j} \theta_{i_1,\dots,i_j}.$$

For every configuration $i_1, \dots, i_j, j = 1, \dots, m$, we define,

$$\theta_{I_j|i_1,\dots,i_{j-1}} = \left\{\theta_{i_j|i_1,\dots,i_{j-1}}\right\}_{i_j=1}^{n_j-1}$$

(for $j = 1$, $\theta_{I_j|i_1,\dots,i_{j-1}} = \{\theta_{i_1}\}_{i_1=1}^{n_1-1}$). Similarly, we define

$$\theta_{i_m,i_1,\dots,i_j} = \sum_{i_{j+1,\dots,}i_{m-1}} \theta_{i_1,\dots,i_m} \quad \text{and} \quad \theta_{i_j|i_m,i_1,\dots,i_{j-1}} = \theta_{i_m,i_1,\dots,i_j} \sum_{i_j} \theta_{i_m,i_1,\dots,i_j}$$

and let

$$\theta_{I_j|i_m,i_1,\dots,i_{j-1}} = \left\{\theta_{i_j|i_m,i_1,\dots,i_{j-1}}\right\}_{i_j=1}^{n_j-1}$$

(for $j = m$, $\theta_{I_j|i_m,i_1,\dots,i_{j-1}} = \{\theta_{i_m}\}_{i_m=1}^{n_m-1}$).

Theorem 3. *Let* $\{\theta_{i_1,\dots,i_m}\}, 1 \le i_j \le n_j, 1 \le j \le m$, *be positive random variables that sum to* 1 *and have a strictly positive pdf* $f_U(\{\theta_{i_1,\dots,i_m}\})$ (*where* m *and* $n_j, j = 1,\dots,m$, *are integers greater than* 1). *Then* $\{\theta_{i_1,\dots,i_m}\}$ *have a Dirichlet distribution—namely the pdf is given by*

$$f_U\left(\{\theta_{i_1,\dots,i_m}\}\right) \propto \prod_{i_1,\dots,i_m} \theta_{i_1,\dots,i_m}^{\alpha_{i_1,\dots,i_m}-1}, \tag{8}$$

where $\alpha_{i_1,\dots,i_m}$ *are positive constants*, *if and only if*

$$\Theta_j = \left\{\theta_{I_j|i_1,\dots,i_{j-1}} | 1 \le i_1 \le n_1, \dots, 1 \le i_{j-1} \le n_{j-1}\right\}$$

are mutually independent (*local parameter independence*), $\{\Theta_j | 1 \le j \le m\}$ *are mutually independent* (*global parameter independence*),

$$\Phi_j = \left\{\theta_{I_j|i_m,i_1,\dots,i_{j-1}} | 1 \le i_m \le n_m, 1 \le i_1 \le n_1, \dots, 1 \le i_{j-1} \le n_{j-1}\right\}$$

are mutually independent and $\{\Phi_j | 1 \le j \le m\}$ *are mutually independent.*

We note that some researchers give Bayesian network structures a causal interpretation [32, 25]. For example, it is common to associate the network $s \to t$ with the statement s causes t, and the network $t \to s$ with the statement t causes s. Under this causal interpretation, define B^c to be the hypothesis that "precisely the independence assertions entailed by B hold in the joint distribution and the edges in B are in the causal direction." Given this definition, it does not follow that $B_1^c = B_2^c$ whenever B_1 and B_2 are Markov equivalent network structures. Nonetheless, it is often reasonable to assume that if B_1 and B_2 are equivalent, then $p(\Theta|B_1^c) = p(\Theta|B_2^c)$, where Θ is the set of all parameters associated with one of the network structures. Under this assumption, our characterization still applies.

4. Discussion.

The independence assumptions made by Theorem 2 can be divided into two parts: $\{\theta_{J|1}, \dots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{I|1}, \dots, \theta_{I|n}\}$ are mutually independent (local parameter independence), and $\theta_{I\cdot}$ is independent of $\{\theta_{J|1}, \dots, \theta_{I|k}\}$ and $\theta_{\cdot J}$ is independent of $\{\theta_{I|1}, \dots, \theta_{i|n}\}$ (global parameter independence). A natural question to ask is whether global parameter independence alone implies a joint Dirichlet pdf for $\{\theta_{ij}\}$.

This question is particularly interesting in light of the analysis of decomposable graphical models given by [6]. Dawid and Lauritzen term a pdf that satisfies global parameter independence a *strong hyper-Markov law*, and show the importance of such laws in the analysis of decomposable graphical models. We now show that the class of strong hyper-Markov laws is larger than the Dirichlet class.

When $n = k = 2$, and using the notation of (5), the new functional equation can be written as follows:

$$f_0(y)g(z,w) = g_0(x)f\left(\frac{yz}{x}, \frac{h(1-z)}{1-x}\right), \tag{9}$$

where $x = yz + (1-y)w$. Note that (5) is obtained from this equation by setting $g(z,w) = g_1(z)g_2(w)$ and $f(t_1,t_2) = f_1(t_1)f_2(t_2)$. These equalities correspond to local parameter independence.

Let f_U be a joint pdf of $\{\theta_{ij}\}$ given by

$$f_U(\{\theta_{ij}\}) = K\left[\prod_{i=1}^{2}\prod_{j=1}^{2}\theta_{ij}^{\alpha_{ij}-1}\right]H\left(\frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}}\right), \tag{10}$$

where K is the normalization constant, α_{ij} are positive constants and H is an arbitrary positive Lebesgue integrable function. That this pdf satisfies global parameter independence can be easily verified. In fact, by solving (9), it can be shown that every positive strong hyper-Markov law can be written in this form (when $n = 2$ and $k = 2$). This solution includes the Dirichlet family as a proper subclass.

Because H is a single function that does not depend on a particular network structure, one can conclude that if local parameter independence holds in *one* network structure, then f_U must still be Dirichlet. Therefore, due to Lemma 1, local parameter independence must hold for *all* network structures. We have proved this claim for two-variable networks, but we believe that it holds for the n-variate case as well. It would be interesting to find specific pdfs of the form given by (10), because such pdfs can be used as priors for the parameters of Bayesian networks while still retaining the advantages of a decomposable prior-to-posterior analysis guaranteed by global parameter independence.

APPENDIX

This Appendix proves Theorem 2. Section A.1 shows that we are allowed to take derivatives of the functional equation which Theorem 2 defines. Section A.2 solves a special case of this functional equation, Section A.3 gives some lemmas needed for the general solution and Sections A.4 and A.5 provide the general solution. Section A.6 uses the general solution to complete the proof of Theorem 2.

A.1. The functional equation. By renaming variable and function names, (4) can be written as follows:

$$\begin{aligned} & f_0(y_1,\dots,y_{n-1})\prod_{j=1}^{n} g_j(z_{1,j},\dots,z_{k-1,j}) \\ & \quad = g_0(x_1,\dots,x_{k-1})\prod_{i=1}^{k} f_i\left(\frac{z_{i1}y_1}{x_i},\dots,\frac{z_{i,n-1}y_{n-1}}{x_i}\right), \end{aligned} \tag{11}$$

where

$$\begin{aligned} x_i &= \sum_{j=1}^{n} z_{ij}y_j, \qquad 1 \le i \le k-1, \\ z_{kj} &= 1 - \sum_{i=1}^{k-1} z_{ij}, \qquad 1 \le j \le n \end{aligned} \tag{12}$$

and where

$$y_n = 1 - \sum_{j=1}^{n-1} y_j, \qquad x_k = 1 - \sum_{i=1}^{k-1} x_i. \tag{13}$$

Note that the free variables in (11) are $y_1,\dots,y_{n-1}$ (y_j replaces $\theta_{\cdot j}$) and z_{ij}, $1 \le i \le k-1$, $1 \le j \le n$ (z_{ij} replaces $\theta_{i|j}$). All other variables which appear in (11) are defined by (12) and (13). Note also that we may consider any y_{j_1} to be a dependent variable instead of y_n as long as $\sum_{j=1}^{n} y_j = 1$, in which case we remain with the same functional equation. Similarly, we may consider x_{i_1} and $z_{i_1 j}$ to be dependent variables instead of x_k and z_{kj}, respectively, as long as $\sum_{i=1}^{k} x_i = 1$ and $\sum_{i=1}^{k} z_{ij} = 1$. These observations are particularly apparent when recalling the probabilistic origin of this equation by which $\{x_i\}_{i=1}^{k}$, for example, are the multinomial parameters associated with a random variable having k states, and no state is distinguished from the other states.

Furthermore, we may consider $x_1,\dots,x_k$ (x_i replaces $\theta_{i\cdot}$) and $w_{ij} = (z_{ij}y_j/x_i)$, $1 \le i \le k$, $1 \le j \le n-1$ (w_{ij} replaces $\theta_{j|i}$) to be free variables and rewrite (11) in terms of these variables. Namely,

$$\begin{aligned} & g_0(x_1,\dots,x_{k-1})\prod_{i=1}^{k} f_i(w_{i,1},\dots,w_{i,n-1}) \\ & \quad = f_0(y_1,\dots,y_{n-1})\prod_{j=1}^{n} g_j\left(\frac{w_{1,j}x_1}{y_j},\dots,\frac{w_{k-1,j}x_{k-1}}{y_j}\right), \end{aligned} \tag{14}$$

where

$$\begin{aligned} y_j &= \sum_{i=1}^{k} w_{ij}x_i, \qquad 1 \le j \le n-1, \\ w_{in} &= 1 - \sum_{j=1}^{n-1} w_{ij}, \qquad 1 \le i \le k \end{aligned} \tag{15}$$

and where x_k and y_n are defined by (13). This symmetric representation of (11) will be used in the derivation of its solution.

We assume that all functions mentioned in (11) originated from pdfs and thus are Lebesgue integrable in their domain. According to Járai's theorems (see Section 2) these assumptions yield that each set of positive functions that solves (11) consists of functions for which any finite-order partial derivative exists for every point in their domain. The importance of this claim is that in order to find all positive Lebesgue integrable functions that satisfy (11), it is permissible to take any derivative at any point in the domain because it exists.

A.2. The bivalued equation. We shall now find all positive Lebesgue integrable solutions of (11) when $k = n = 2$. This derivation is different from the general derivation which is given in the next sections.

When $k = n = 2$, (11) reduces to

$$f_0(y)g_1(z)g_2(w) = g_0(x)f_1\left(\frac{yz}{x}\right)f_2\left(\frac{y(1-z)}{1-x}\right), \tag{16}$$

where

$$x = yz + (1-y)w. \tag{17}$$

Let

$$\hat{f}_i'(t) = \frac{d}{dt}\ln f_i(t), \tag{18}$$

and

$$\hat{g}_i'(t) = \frac{d}{dt}\ln g_i(t). \tag{19}$$

Taking the logarithm and then a derivative once wrt y, once wrt z and once wrt w of (16) yields the following three equations:

$$\begin{aligned}&\hat{f}_0'(y) - (z-w)\hat{g}_0'(x)\\ &\quad= \frac{zw}{x^2}\hat{f}_1'\left(\frac{yz}{x}\right) + \frac{(1-z)(1-w)}{(1-x)^2}\hat{f}_2'\left(\frac{y(1-z)}{1-x}\right),\end{aligned} \tag{20}$$

$$\begin{aligned}&\hat{g}_1'(z) - y\hat{g}_0'(x)\\ &\quad= \frac{yw(1-y)}{x^2}\hat{f}_1'\left(\frac{yz}{x}\right) - \frac{(1-w)(1-y)y}{(1-x)^2}\hat{f}_2'\left(\frac{y(1-z)}{1-x}\right),\end{aligned} \tag{21}$$

$$\begin{aligned}&\hat{g}_2'(w) - (1-y)\hat{g}_0'(x)\\ &\quad= -\frac{yz(1-y)}{x^2}\hat{f}_1'\left(\frac{yz}{x}\right) + \frac{y(1-z)(1-y)}{(1-x)^2}\hat{f}_2'\left(\frac{y(1-z)}{1-x}\right).\end{aligned} \tag{22}$$

Solving $\hat{f}_1'(yz/x)$ and $\hat{f}_2'(y(1-z)/(1-x))$ from (21) and (22), plugging the result back into (20) and collecting all the terms involving $\hat{g}_0'(x)$, $\hat{g}_1'(z)$, $\hat{g}_2'(w)$

and $\hat{f}_0'(y)$ yields

$$\begin{aligned} &h(y,z,w)\hat{g}_0'(x) \\ &\quad = z(1-z)\hat{g}_1'(z) + w(1-w)\hat{g}_2'(w) - y(1-y)(w-z)\hat{f}_0'(y), \end{aligned} \tag{23}$$

where

$$h(y,z,w) = y(1-y)(w-z)^2 + yz(1-z) + (1-y)(1-w)w.$$

Taking a derivative wrt z of (23) and multiplying the result by $1-y$, and similarly, taking a derivative wrt w of (23) and multiplying the result by y yields, after subtracting the two equations,

$$\begin{aligned} &[(1-y)h_z(y,z,w) - yh_w(y,z,w)]\hat{g}_0'(x) \\ &\quad = (1-y)[(1-2z)\hat{g}_1'(z) + z(1-z)\hat{g}_1''(z)] \\ &\qquad + (1-y)[y(1-y)\hat{f}_0'(y)] \\ &\qquad - y[(1-2w)\hat{g}_2'(w) + w(1-w)\hat{g}_2''(w) - y(1-y)\hat{f}_0'(y)], \end{aligned} \tag{24}$$

where h_z and h_w are the partial derivatives of h wrt z and w, respectively. But we also have

$$(1-y)h_z(y,z,w) - yh_w(y,z,w) \equiv 0$$

and therefore (24) yields

$$\begin{aligned} &(1-2w)\hat{g}_2'(w) + w(1-w)\hat{g}_2''(w) \\ &\quad = \frac{1-y}{y}[(1-2z)\hat{g}_1'(z) + z(1-z)\hat{g}_1''(z)] + (1-y)\hat{f}_0'(y). \end{aligned} \tag{25}$$

Because w does not appear in the right-hand side of this equation, we get

$$(1-2w)\hat{g}_2'(w) + w(1-w)\hat{g}_2''(w) = c_1, \tag{26}$$

where c_1 is an arbitrary constant. Equation 26 is a first-order linear differential equation, the general solution of which is given by

$$\hat{g}_2'(w) = \frac{b}{w(1-w)} - \frac{c_1}{2}\frac{1-2w}{w(1-w)},$$

where b is an arbitrary constant and $b/w(1-w)$ is the homogeneous solution. Thus,

$$\hat{g}_2'(w) = \frac{\alpha}{w} - \frac{\beta}{1-w},$$

where α and β are arbitrary constants defined by $\alpha = b - c_1/2$ and $\beta = -(b + 3c_1/2)$. Hence, $g_2(w) = cw^{\alpha}(1-w)^{\beta}$ where c is a third arbitrary constant.

From (25) we also get

$$(1-2z)\hat{g}_1'(z) + z(1-z)\hat{g}_1''(z) = \frac{c_1 y}{1-y} + u\hat{f}_0'(y).$$

Hence both sides are equal to a constant, say c_2. Consequently,

$$\hat{f}_0'(y) = \frac{c_2}{y} - \frac{c_1}{1-y}$$

and

$$\hat{g}_1'(z) = \frac{\alpha'}{z} - \frac{\beta'}{1-z}.$$

Consequently, $f_0(y)$, $g_1(z)$ and $g_2(w)$ all have the Dirichlet functional form and each function depends on three constants.

A.3. Preliminary lemmas. We now provide several lemmas that are needed for the derivation of the general solution of (11).

LEMMA A.1. *The general solution of the following partial differential equation for* $f(x_1, \dots, x_n)$,

$$f + x_i f_{x_i} + x_j f_{x_j} = 0 \tag{27}$$

in the domain $(0, \infty)^m$, *is given by*

$$f(x_1, \dots, x_n) = \frac{1}{x_i} h\left(\frac{x_i}{x_j}, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_n\right) \tag{28}$$

or, equivalently, by

$$f(x_1, \dots, x_n) = \frac{1}{x_j} g\left(\frac{x_i}{x_j}, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_n\right), \tag{29}$$

where h and g are arbitrary differentiable functions having $n - 1$ *arguments.*

PROOF. Let $s = x_i$ and $t = x_i/x_j$. Thus, $f_{x_i} = f_s + (t/s)f_t$, $f_{x_j} = -(t^2/s)f_t$. Hence, after a change of variables, the differential equation becomes

$$f + sf_s = 0$$

and therefore $f = (1/s)h(t, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_n)$. By changing the roles of x_i and x_j in this derivation, we get the other form of f. □

LEMMA A.2. *The general solution of the following partial differential equation for* $f(x_1, \dots, x_n)$,

$$f_{x_i} - f_{x_j} = \frac{\alpha}{x_i} + \frac{\beta}{x_j}, \tag{30}$$

is given by

$$\begin{aligned} &f(x_1, \dots, x_n) \\ &\quad = \alpha \ln x_i - \beta \ln x_j \\ &\qquad + h(x_i + x_j, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_n), \end{aligned} \tag{31}$$

where h is an arbitrary differentiable function having $n - 1$ *arguments.*

PROOF. Let $s = x_i + x_j$ and $t = x_i - x_j$. Thus, $f_{x_i} = f_s + f_t, f_{x_j} = f_s - f_t$. Hence, after a change of variables, the differential equation becomes

$$f_t = \frac{\alpha}{s+t} + \frac{\beta}{s-t}.$$

Integrating wrt t and changing back to the original variables yields the desired solution. □

LEMMA A.3. *Let $f(x_1, \dots, x_n)$ be a twice-differentiable function. If for all $1 \le i < j \le n$,*

$$f(x_1, \dots, x_n) = a_i \ln x_i + a_j \ln x_j + f_{ij}(x_i + x_j, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_n),$$

where f_{ij} are arbitrary twice-differentiable functions having $n - 1$ arguments, then

$$f(x_1, \dots, x_n) = g\left(\sum_{i=1}^{n} x_i\right) + \sum_{i=1}^{n} a_i \ln x_i, \tag{32}$$

where g is an arbitrary twice-differentiable function.

Proof can be found in [9].

A.4. The general solution. We now solve (11) for any n and k. First we assume n and k are strictly greater than 2. We use the following notation:

$$\begin{aligned} &g_l(t_1, \dots, t_{k-1})_i \\ &\quad= \frac{\partial}{\partial t_i} \ln g_l(t_1, \dots, t_{k-1}), \qquad 1 \le i \le k-1, 0 \le l \le n, \\ &g_l(t_1, \dots, t_{k-1})_{ij} \\ &\quad= \frac{\partial}{\partial t_i}\frac{\partial}{\partial t_j} \ln g_l(t_1, \dots, t_{k-1}), \qquad 1 \le i, j \le k-1, 0 \le l \le n, \\ &f_l(t_1, \dots, t_{n-1})_i \\ &\quad= \frac{\partial}{\partial t_i} \ln f_l(t_1, \dots, t_{n-1}), \qquad 1 \le i \le n-1, 0 \le l \le k, \\ &f_l(t_1, \dots, t_{n-1})_{ij} \\ &\quad= \frac{\partial}{\partial t_i}\frac{\partial}{\partial t_j} \ln f_l(t_1, \dots, t_{n-1}), \qquad 1 \le i, j \le n-1, 0 \le l \le k. \end{aligned} \tag{33}$$

Also we use the following notation:

$$\begin{aligned} &X = (x_1, \dots, x_{k-1}), \qquad Z_j = (z_{1,j}, \dots, z_{k-1,j}), \\ &Y = (y_1, \dots, y_{n-1}), \qquad W_i = \left(\frac{z_{i1} y_1}{x_i}, \dots, \frac{z_{i,n-1} y_{n-1}}{x_i}\right). \end{aligned} \tag{34}$$

Thus, for example, $g_j(Z_j)$ stands for $g_j(z_{1,j}, \dots, z_{k-1,j})$.

By taking the logarithm and then a derivative wrt z_{ij} $(1 \le i \le k-1, 1 \le j \le n-1)$ of (11), we get

$$g_j(Z_j)_i = y_j g_0(X)_i + y_j\left[\sum_{l=1}^{n-1} f_i(W_i)_l\left[-\frac{z_{il}y_l}{x_i^2}\right] + \frac{1}{x_i}f_i(W_i)_j\right] \tag{35}$$
$$+ y_j\left[\sum_{l=1}^{n-1} f_k(W_k)_l\left[\frac{z_{kl}y_l}{x_k^2}\right] - \frac{1}{x_k}f_k(W_k)_j\right].$$

By setting $i = i_1$ and $i = i_2$, $1 \le i_1 < i_2 \le k-1$ $(k \ge 3)$ in (35), subtracting the resulting two equations and dividing by y_j, we get

$$\frac{1}{y_j}\left[g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}\right]$$
$$= \left[g_0(X)_{i_1} - g_0(X)_{i_2}\right] \tag{36}$$
$$+ \sum_{l=1}^{n-1}\left[f_{i_2}(W_{i_2})_l\left[\frac{z_{i_2 l}y_l}{x_{i_2}^2}\right] - f_{i_1}(W_{i_1})_l\left[\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right]\right]$$
$$+ \frac{1}{x_{i_1}}f_{i_1}(W_{i_1})_j - \frac{1}{x_{i_2}}f_{i_2}(W_{i_2})_j.$$

Now taking the logarithm and then a derivative wrt z_{in} $(1 \le i \le k-1)$ of (11) yields

$$g_n(Z_n)_i = y_n g_0(X)_i + y_n\left[\sum_{l=1}^{n-1} f_i(W_i)_l\left[-\frac{z_{il}y_l}{x_i^2}\right]\right] \tag{37}$$
$$+ y_n\left[\sum_{l=1}^{n-1} f_k(W_k)_l\left[\frac{z_{kl}y_l}{x_k^2}\right]\right].$$

Similarly, by setting $i = i_1$ and $i = i_2$, $1 \le i_1 < i_2 \le k-1$ in (37), subtracting the resulting two equations and dividing by y_n, we get

$$\frac{1}{y_n}\left[g_n(Z_n)_{i_1} - g_n(Z_n)_{i_2}\right]$$
$$= \left[g_0(X)_{i_1} - g_0(X)_{i_2}\right] \tag{38}$$
$$+ \sum_{l=1}^{n-1}\left[f_{i_2}(W_{i_2})_l\left[\frac{z_{i_2 l}y_l}{x_{i_2}^2}\right] - f_{i_1}(W_{i_1})_l\left[\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right]\right].$$

Subtracting (38) from (36) and setting $j = j_1$ yields

$$\frac{1}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1} - g_{j_1}(Z_{j_1})_{i_2}\right] - \frac{1}{y_n}\left[g_n(Z_n)_{i_1} - g_n(Z_n)_{i_2}\right] \tag{39}$$
$$= \frac{1}{x_{i_1}}f_{i_1}(W_{i_1})_{j_1} - \frac{1}{x_{i_2}}f_{i_2}(W_{i_2})_{j_1},$$

where $1 \le i_1 < i_2 \le k-1, 1 \le j_1 \le n-1$.

Now we take a derivative wrt $z_{i_1j_1}$ of (39) and obtain

$$\begin{aligned}&\frac{1}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1i_1} - g_{j_1}(Z_{j_1})_{i_2i_1}\right]\\ &\quad = -\frac{y_{j_1}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1} + \frac{y_{j_1}}{x_{i_1}}\sum_{l=l}^{n-1} f_{i_1}(W_{i_1})_{j_1 l}\left[-\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right] + \frac{y_{j_1}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1j_1}.\end{aligned} \tag{40}$$

Similarly, we take a derivative wrt z_{i_1n} of (39) and obtain

$$\begin{aligned}&-\frac{1}{y_n}\left[g_n(Z_n)_{i_1i_1} - g_n(Z_n)_{i_2i_1}\right]\\ &\quad = -\frac{y_n}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1} + \frac{y_n}{x_{i_1}}\sum_{l=1}^{n-1} f_{i_1}(W_{i_1})_{j_1 l}\left[-\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right].\end{aligned} \tag{41}$$

Equations (40) and (41) yield

$$\begin{aligned}&\frac{1}{y_{j_1}^2}\left[g_{j_1}(Z_{j_1})_{i_1i_1} - g_{j_1}(Z_{j_1})_{i_2i_1}\right] + \frac{1}{y_n^2}\left[g_n(Z_n)_{i_1i_1} - g_n(Z_n)_{i_2i_1}\right]\\ &\quad = \frac{1}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1j_1}.\end{aligned} \tag{42}$$

Now we take a derivative wrt $z_{i_1j_2}$ of (39) where $1 \le j_2 \le n-1, j_2 \ne j_1$ ($n \ge 3$), and obtain

$$0 = -\frac{y_{j_2}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1} + \frac{y_{j_2}}{x_{i_1}}\sum_{l=1}^{n-1} f_{i_1}(W_{i_1})_{j_1 l}\left[-\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right] + \frac{y_{j_2}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1j_2}. \tag{43}$$

Equations (41) and (43) yield ($j_1 \ne j_2$)

$$\frac{1}{y_n^2}\left[g_n(Z_n)_{i_1i_1} - g_n(Z_n)_{i_2i_1}\right] = \frac{1}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1j_2}. \tag{44}$$

Putting (42) and (44) into (43) and recalling [from (12)] that

$$z_{i_1n}y_n = x_{i_1} - \sum_{l=1}^{n-1} z_{i_1l}y_l,$$

we get

$$\begin{aligned}\frac{1}{x_{i_1}} f_{i_1}(W_{i_1})_{j_1} &= -\frac{z_{i_1j_1}}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1i_1} - g_{j_1}(Z_{j_1})_{i_2i_1}\right]\\ &\quad + \frac{z_{i_1n}}{y_n}\left[g_n(Z_n)_{i_1i_1} - g_n(Z_n)_{i_2i_1}\right].\end{aligned} \tag{45}$$

Similarly, we derive an analogue to (40) by taking a derivative wrt $z_{i_2 j_1}$ (instead of wrt $z_{i_1 j_1}$) of (39), follow the same steps up to (45) and get

$$\begin{aligned}\frac{1}{x_{i_2}} f_{i_2}(W_{i_2})_{j_1} = &-\frac{z_{i_2 j_1}}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1 i_2} - g_{j_1}(Z_{j_1})_{i_2 i_2}\right] \\ &+\frac{z_{i_2 n}}{y_n}\left[g_n(Z_n)_{i_1 i_2} - g_n(Z_n)_{i_2 i_2}\right].\end{aligned} \tag{46}$$

Plugging (45) and (46) into (39) and collecting all terms involving y_n on one side and all terms not involving y_n on the other side implies that each side is equal to a constant, say c. Namely,

$$\begin{aligned}\frac{1}{y_j}\left[g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}\right] + \frac{z_{i_1 j}}{y_j}\left[g_j(Z_j)_{i_1 i_1} - g_j(Z_j)_{i_2 i_1}\right] \\ + \frac{z_{i_2 j}}{y_j}\left[g_j(Z_j)_{i_1 i_2} - g_j(Z_j)_{i_2 i_2}\right] = c,\end{aligned} \tag{47}$$

where $1 \le j \le n$.

This equation holds for every value of y_j and therefore $c = 0$. Thus we obtain

$$\begin{aligned}\left[g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}\right] + z_{i_1 j}\left[g_j(Z_j)_{i_1 i_1} - g_j(Z_j)_{i_2 i_1}\right] \\ + z_{i_2 j}\left[g_j(Z_j)_{i_1 i_2} - g_j(Z_j)_{i_2 i_2}\right] = 0.\end{aligned} \tag{48}$$

Let $h(Z_j) = g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}$. Thus (48) can be written as follows:

$$h + z_{i_1 j}\frac{\partial h}{\partial z_{i_1 j}} + z_{i_2 j}\frac{\partial h}{\partial z_{i_2 j}} = 0. \tag{49}$$

Lemma A.1 provides the general solution for h and thus,

$$h(Z_j) = g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2} = \frac{1}{z_{i_1 j}}\tilde{g}_j\left(\frac{z_{i_1 j}}{z_{i_2 j}}, Z_{i_1 i_2, j}\right), \tag{50}$$

where

$$Z_{i_1 i_2, j} = (z_{1j}, \dots, z_{i_1 - 1, j}, z_{i_1 + 1, j}, \dots, z_{i_2 - 1, j}, z_{i_2 + 1, j}, \dots, z_{k-1, j})$$

and where $\tilde{g}_j$ is an arbitrary function having one argument less than g_j, or

$$g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2} = \frac{1}{z_{i_2 j}}\tilde{g}_j\left(\frac{z_{i_1 j}}{z_{i_2 j}}, Z_{i_1 i_2, j}\right), \tag{51}$$

where again $\tilde{g}_j$ is an arbitrary function having one argument less than g_j. Similarly, because f_i and g_j play a symmetric role in (11) as shown by (14) and hence have the same form, we get

$$f_i(W_i)_{j_1} - f_i(W_i)_{j_2} = \frac{x_i}{z_{i j_1} y_{j_1}}\tilde{f}_i\left(\frac{z_{i j_1} y_{j_1}}{z_{i j_2} y_{j_2}}, W_{j_1 j_2, i}\right), \tag{52}$$

where

$$W_{j_1j_2,i} = \left(\frac{z_{i_1}y_1}{x_i}, \dots, \frac{z_{i,j_1-1}y_{j_1-1}}{x_i}, \frac{z_{i,j_1+1}y_{j_1+1}}{x_i}, \dots, \right.$$
$$\left. \frac{z_{i,j_2-1}y_{j_2-1}}{x_i}, \frac{z_{i,j_2+1}y_{j_2+1}}{x_i}, \dots, \frac{z_{in}y_n}{x_i} \right)$$

or

$$f_i(W_i)_{j_1} - f_i(W_i)_{j_2} = \frac{x_i}{z_{ij_2}y_{j_2}} \tilde{f}_i\left(\frac{z_{ij_1}y_{j_1}}{z_{ij_2}y_{j_2}}, W_{j_1j_2,i} \right). \tag{53}$$

Now, by setting $j = j_1$ and $j = j_2$ in (39) and subtracting the resulting equations, we get

$$\begin{aligned} &\frac{1}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1} - g_{j_1}(Z_{j_1})_{i_2} \right] - \frac{1}{y_{j_2}}\left[g_{j_2}(Z_{j_2})_{i_1} - g_{j_2}(Z_{j_2})_{i_2} \right] \\ &\quad = \frac{1}{x_{i_1}}\left[f_{i_1}(W_{i_1})j_1 - f_{i_1}(W_{i_1})_{j_2} \right] - \frac{1}{x_{i_2}}\left[f_{i_2}(W_{i_2})_{j_1} - f_{i_2}(W_{i_2})_{j_2} \right]. \end{aligned} \tag{54}$$

Plugging (50) through (53) into (54) yields

$$\begin{aligned} &\frac{1}{z_{i_1j_1}y_{j_1}} \tilde{g}_{j_1}\left(\frac{z_{i_1j_1}}{z_{i_2j_1}}, Z_{i_1i_2,j_1} \right) - \frac{1}{z_{i_2j_2}y_{j_2}} \tilde{g}_{j_2}\left(\frac{z_{i_1j_2}}{z_{i_2j_2}}, Z_{i_1i_2,j_2} \right) \\ &\quad = \frac{1}{z_{i_1j_1}y_{j_1}} \tilde{f}_{i_1}\left(\frac{z_{i_1j_1}y_{j_1}}{z_{i_1j_2}y_{j_2}}, W_{j_1j_2,i_1} \right) - \frac{1}{z_{i_2j_2}y_{j_2}} \tilde{f}_{i_2}\left(\frac{z_{i_2j_1}y_{j_1}}{z_{i_2j_2}y_{j_2}}, W_{j_1j_2,i_2} \right) \end{aligned} \tag{55}$$

Note that the variables in $Z_{i_1i_2,j_1}$ do not appear elsewhere in this equation. Therefore, $\tilde{g}_{j_1}$ is only a function of its first argument. Similarly, $\tilde{g}_{j_2}$, $\tilde{f}_{i_1}$ and $\tilde{f}_{i_2}$ are only functions of their first argument. Thus (55) can be rewritten as follows:

$$\begin{aligned} &\frac{1}{z_{i_1j_1}y_{j_1}} \tilde{\tilde{g}}_{j_1}\left(\frac{z_{i_1j_1}}{z_{i_2j_1}} \right) - \frac{1}{z_{i_2j_2}y_{j_2}} \tilde{\tilde{g}}_{j_2}\left(\frac{z_{i_1j_2}}{z_{i_2j_2}} \right) \\ &\quad = \frac{1}{z_{i_1j_1}y_{j_1}} \tilde{\tilde{f}}_{i_1}\left(\frac{z_{i_1j_1}y_{j_1}}{z_{i_1j_2}y_{j_2}} \right) - \frac{1}{z_{i_2j_2}y_{j_2}} \tilde{\tilde{f}}_{i_2}\left(\frac{z_{i_2j_1}y_{j_1}}{z_{i_2j_2}y_{j_2}} \right). \end{aligned} \tag{56}$$

Let $x = z_{i_1j_1}y_{j_1}$, $y = z_{i_2j_1}y_{j_1}$, $z = z_{i_1j_2}y_{j_2}$ and $w = z_{i_2j_2}y_{j_2}$ in (56). Then

$$\frac{1}{x}\left[\tilde{\tilde{g}}_{j_1}\left(\frac{x}{y} \right) - \tilde{\tilde{f}}_{i_1}\left(\frac{x}{z} \right) \right] = \frac{1}{w}\left[\tilde{\tilde{g}}_{j_2}\left(\frac{z}{x} \right) - \tilde{\tilde{f}}_{i_2}\left(\frac{y}{w} \right) \right]. \tag{57}$$

By taking a derivative wrt y of (57), we get

$$\frac{\tilde{\tilde{g}}'_{j_1}(x/y)}{y^2} = \frac{\tilde{\tilde{f}}'_{i_2}(y/w)}{w^2}. \tag{58}$$

Setting $y = w$, we see that $\tilde{\tilde{g}}'_{j_1}(t) = \beta_{j_1}$ and $\tilde{\tilde{g}}_{j_1}(t) = \beta_{j_1} t + \alpha_{j_1}$, where α_{j_1} and β_{j_1} are constants. Plugging this result into (50) yields

$$g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2} = \frac{\alpha}{z_{i_1 j}} + \frac{\beta}{z_{i_2 j}}, \tag{59}$$

where $1 \le i_1 < i_2 \le k - 1$.

Equation (59) is a first-order partial differential equation, the general solution of which is given by Lemma A.2. Consequently, due to (33), we get

$$\begin{aligned} &g_j(t_1, \dots, t_{k-1}) \\ &\quad = t_{i_1}^{\alpha_{i_1 j}} t_{i_2}^{\alpha_{i_2 j}} g_j\bigl(t_{i_1} + t_{i_2}, t_1, \dots, t_{i_1 - 1}, t_{i_1 + 1}, \dots, t_{i_2 - 1}, t_{i_2 + 1}, \dots, t_{k-1}\bigr). \end{aligned} \tag{60}$$

Now, due to Lemma A.3 we have

$$g_j(t_1, \dots, t_{k-1}) = \left[\prod_{i=1}^{k-1} t_i^{\alpha_{ij}}\right] G_j\left(\sum_{i=1}^{k-1} t_i\right). \tag{61}$$

Similarly,

$$f_i(t_1, \dots, t_{n-1}) = \left[\prod_{j=1}^{n-1} t_j^{\beta_{ij}}\right] F_i\left(\sum_{j=1}^{n-1} t_j\right), \tag{62}$$

which is obtained by repeating the derivation starting at (14) rather then at (11). Note that we have almost derived the Dirichlet functional form. It remains to derive the form of the functions F_i and G_j.

In (11) let $z_{1j} = z_{2j} = \cdots = z_{kj}$ for $1 \le j \le n$. Thus, according to (12), $z_{ij} = x_i$. Consequently, we get

$$\begin{aligned} &f_0(y_1, \dots, y_{n-1}) \prod_{j=1}^{n} g_j(x_1, \dots, x_{k-1}) \\ &\qquad = g_0(x_1, \dots, x_{k-1}) \prod_{i=1}^{k} f_i(y_1, \dots, y_{n-1}). \end{aligned} \tag{63}$$

Equations (11) and (63) yield

$$\prod_{j=1}^{n} \frac{g_j(z_{1,j}, \dots, z_{k-1,n})}{g_j(x_1, \dots, x_{k-1})} = \prod_{i=1}^{k} \frac{f_i(z_{i1} y_1/x_i, \dots, z_{i,n-1} y_{n-1}/x_i)}{f_i(y_1, \dots, y_{n-1})}. \tag{64}$$

Plugging (61) and (62) into (64), we get

$$\begin{aligned} &\left[\prod_{j=1}^{n} \prod_{i=1}^{k-1} \left(\frac{z_{ij}}{x_i}\right)^{\alpha_{ij}}\right] \left[\prod_{j=1}^{n} \frac{G_j\bigl(\sum_{i=1}^{k-1} z_{ij}\bigr)}{G_j\bigl(\sum_{i=1}^{k-1} x_i\bigr)}\right] \\ &\qquad = \left[\prod_{i=1}^{k} \prod_{j=1}^{n-1} \left(\frac{z_{ij}}{x_i}\right)^{\beta_{ij}}\right] \left[\prod_{i=1}^{k} \frac{F_i\bigl(\sum_{j=1}^{n-1} (z_{ij} y_j)/x_i\bigr)}{F_i\bigl(\sum_{j=1}^{n-1} y_j\bigr)}\right]. \end{aligned} \tag{65}$$

Thus, using $z_{kj} = 1 - \sum_{i=1}^{k-1} z_{ij}$ [by (12)],

$$\left[\prod_{j=1}^{n-1}\prod_{i=1}^{k-1}\left(\frac{z_{ij}}{x_i}\right)^{c_{ij}}\right]\left[\prod_{j=1}^{n}\frac{\tilde{G}_j\left(\sum_{i=1}^{k-1} z_{ij}\right)}{\tilde{G}_j\left(\sum_{i=1}^{k-1} x_i\right)}\right] = \prod_{i=1}^{k}\frac{\tilde{F}_i\left(\sum_{j=1}^{n-1}\left(z_{ij}y_j/x_i\right)\right)}{\tilde{F}_i\left(\sum_{j=1}^{n-1} y_j\right)}, \tag{66}$$

where for $1 \le i \le k-1$ and $1 \le j \le n-1$, $c_{ij} = \alpha_{ij} - \beta_{ij}$,

$$\tilde{F}_i(t) = (1-t)^{-\alpha_{in}} F_i(t), \qquad \tilde{G}_j(t) = (1-t)^{-\beta_{kj}} G_j(t)$$

and where $\tilde{F}_k(t) = F_k(t)$ and $\tilde{G}_n(t) = G_n(t)$. We will show that $\tilde{F}_i(t)$, $i = 1, \dots, k-1$, are constants. Consequently, due to (62), f_i has a Dirichlet functional form. That the function f_k also has a Dirichlet functional form can be obtained by choosing z_{1j} as a dependent variable defined by $z_{1j} = 1 - \sum_{i=2}^{k} z_{ij}$ instead of z_{kj} as defined by (12) and repeating the same arguments. By symmetric arguments, each g_j also has a Dirichlet functional form.

Let $y_j = 1/n$, for all j, $1 \le j \le n$ and $z_{ij} = 1/k$ for all i and j, $1 \le i \le k$, $1 \le j \le n-1$. Hence, the only free variables remaining in (66) are z_{in} where $1 \le i \le k-1$. Note that $x_i = \sum_{j=1}^{n} z_{ij}y_j = (n-1)/kn + (1/n)z_{in}$, $1 \le i \le k-1$, and so $\tilde{G}_j(\sum_{i=1}^{k-1} x_i)$ is a function of $\sum_{i=1}^{k-1} z_{in}$. Also $\tilde{G}_j(\sum_{i=1}^{k-1} z_{ij})$ is a constant for $1 \le j \le n-1$ and a function of $\sum_{i=1}^{k-1} z_{in}$ for $j = n$. Consequently, (66) becomes

$$f\left(\sum_{i=1}^{k-1} z_{in}\right) = \prod_{i=1}^{k-1}\tilde{F}_i\left(\frac{c}{c + dz_{in}}\right)\left[\frac{c + dz_{in}}{c}\right]^{a_i}, \tag{67}$$

where $c = (n-1)/kn$, $d = 1/n$ and $a_i = \sum_{j=1}^{n-1} c_{ij}$. Note that $z_{kn} = 1 - \sum_{i=1}^{k-1} z_{in}$ and so the kth term on the right-hand side of (66) is absorbed, along with some constants, into the definition of f in (67).

Let $t_i = c/(c + dz_{in})$; $z_{in} = (c/d)((1-t_i)/t_i)$. Taking the logarithm of (67), we get

$$\hat{f}\left(\frac{c}{d}\sum_{i=1}^{k-1}\frac{1-t_i}{t_i}\right) = \sum_{i=1}^{k-1}\ln t_i^{-a_i}\tilde{F}_i(t_i). \tag{68}$$

Taking a derivative wrt t_{i_1}, $1 \le i_1 \le k-1$, we get

$$-\frac{c}{dt_{i_1}^2}\hat{f}'\left(\frac{c}{d}\sum_{i=1}^{k-1}\frac{1-t_i}{t_i}\right) = \left[\ln t_i^{-a_{i_1}}\tilde{F}_{i_1}(t_{i_1})\right]'. \tag{69}$$

Thus, $\hat{f}'((c/d)\sum_{i=1}^{k-1}((1-t_i)/t_i))$ must be a constant. Hence, by integrating (69),

$$\tilde{F}_i(t) = c_i t^{a_i} e^{K/t}, \qquad 1 \le i \le k-1, \tag{70}$$

where K is a constant not depending on i.

To complete the derivation, we substitute (70) into (66), and let $y_j = 1/n$, for $1 \le j \le n$ and $z_{ij} = 1/k$ except z_{i1}, $1 \le i \le k-1$ which remain free variables. Consequently, we get

$$g\left(\sum_{i=1}^{k-1} z_{i_1}\right) = \prod_{i=1}^{k-1} \frac{(z_{i1}+w_0)^{a_i}}{z_{i1}^{c_{i1}}} \exp\left(K \sum_{k-1}^{i=1} \frac{1}{z_{i1}+w_0}\right),$$

where $w_0 = (n-2)/k$. Therefore, $K = 0, a_i = 0$ and $\tilde{F}_i$ is a constant as claimed.

Thus,

$$f_i(t_1,\dots,t_{n-1}) = k_i \left[\prod_{j=1}^{n-1} t_j^{\beta_j}\right]\left(1 - \sum_{j=1}^{n-1} t_j\right)^{\beta_k}, \tag{71}$$

$$g_j(t_1,\dots,t_{k-1}) = c_j \left[\prod_{i=1}^{k-1} t_i^{\alpha_i}\right]\left(1 - \sum_{i=1}^{k-1} t_i\right)^{\alpha_k}. \tag{72}$$

A.5. Special cases. We now solve (11) when $n = 2$ and $k \ge 3$. This proof follows the general lines presented in Section A.4 but circumvents the applications of the fact $n \ge 3$ assumed in Section A.4. When $k = 3$ and $n \ge 3$, a similar derivation can be obtained, as implied by the symmetric roles of n and k in (4).

Note that up to (42) the derivation is valid when $n = 2$. Furthermore, note that the sum in (41) consists now of one term, where $l = j_1 = 1$. Thus, (41) and (42) yield, using $x_i = z_{ij_1} y_{j_1} + z_{in} y_n$ $(n = 2, j_1 = 1)$,

$$\begin{aligned}\frac{f_{i_1}(W_{i_1})_{j_1}}{x_{i_1}} &= \frac{z_{i_1 n}}{y_n}\left[g_n(Z_n)_{i_1 i_1} - g_n(Z_n)_{i_2 i_1}\right] \\ &\quad - \frac{z_{i_1 j_1}}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1 i_1} - g_{j_1}(Z_{j_1})_{i_2 i_1}\right].\end{aligned} \tag{73}$$

Similarly,

$$\begin{aligned}\frac{f_{i_2}(W_{i_2})_{j_1}}{x_{i_2}} &= \frac{z_{i_2 n}}{y_n}\left[g_n(Z_n)_{i_1 i_2} - g_n(Z_n)_{i_2 i_2}\right] \\ &\quad - \frac{z_{i_2 j_1}}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1 i_2} - g_{j_1}(Z_{j_1})_{i_2 i_2}\right],\end{aligned} \tag{74}$$

which is obtained by taking a derivative wrt $z_{i_2 j_1}$ of (39) (instead of wrt $z_{i_1 j_1}$) and repeating the derivation up to (42).

Plugging (73) and (74) into (39) and collecting all terms involving y_n on one side and all terms not involving y_n on the other side implies that each side is equal to a constant, say c, namely, we obtain the partial differential equation for $g_j(Z_j)$, $1 \le j \le n$, given by (47). Consequently, as given by (50) and because $n = 2$,

$$g_{j_1}(Z_{j_1})_{i_1} - g_{j_1}(Z_{j_1})_{i_2} = \frac{1}{z_{i_1 j_1}} \hat{g}_{j_1}\left(\frac{z_{i_1 j_1}}{z_{i_2 j_1}}\right) \tag{75}$$

and

$$g_{j_2}(Z_{j_2})_{i_1} - g_{j_2}(Z_{j_2})_{i_2} = \frac{1}{z_{i_2 j_2}} \hat{g}_{j_2}\left(\frac{z_{i_1 j_2}}{z_{i_2 j_2}} \right). \tag{76}$$

Also, when $n = 2$, we have $x_i = z_{ij_1} y_{j_1} + z_{in} y_n$, and hence,

$$\frac{1}{x_{i_1}} f_{i_1}(W_{i_1})_{j_1} = \frac{1}{x_{i_1}} f_{i_1}\left(\frac{z_{i_1 j_1} y_{j_1}}{x_{i_1}} \right)_{j_1} = \frac{1}{z_{i_1 j_1} y_{j_1}} \hat{f}_{i_1}\left(\frac{z_{i_1 j_1} y_{j_1}}{z_{i_1 n} y_n} \right), \tag{77}$$

$$\frac{1}{x_{i_2}} f_{i_2}(W_{i_2})_{j_1} = \frac{1}{x_{i_2}} f_{i_2}\left(\frac{z_{i_2 j_1} y_{j_1}}{x_{i_2}} \right)_{j_1} = \frac{1}{z_{i_2 j_1} y_{j_1}} \hat{f}_{i_2}\left(\frac{z_{i_2 j_1} y_{j_1}}{z_{i_2 n} y_n} \right). \tag{78}$$

Plugging (75) and (78) into (39) yields

$$\begin{aligned} &\frac{1}{z_{i_1 j_1} y_{j_1}} \hat{g}_{j_1}\left(\frac{z_{i_1 j_1}}{z_{i_2 j_1}} \right) - \frac{1}{z_{i_2 n} y_n} \hat{g}_n\left(\frac{z_{i_1 n}}{z_{i_2 n}} \right) \\ &\quad = \frac{1}{z_{i_1 j_1} y_{j_1}} \hat{f}_{i_1}\left(\frac{z_{i_1 j_1} y_{j_1}}{z_{i_1 n} y_n} \right) - \frac{1}{z_{i_2 n} y_n} \hat{f}_{i_2}\left(\frac{z_{i_2 j_1} y_{j_1}}{z_{i_2 n} y_n} \right). \end{aligned} \tag{79}$$

This equation parallels (56) where j_2 is replaced by n and can be solved in the same way. Thus (61) is obtained. Equation (62), on the other hand, needs no proof when $n = 2$ because an arbitrary function $f(x)$ defined on $(0, 1)$ can always be written as $f(x) = x^{\alpha} g(x)$ where $g(x) = x^{-\alpha} f(x)$. The remainder of the derivation follows Section A.4 closely.

A.6. The joint density. In previous sections we have shown that, under the assumptions made by Theorem 2, the densities $f_I(\theta_{I\cdot})$ and $f_{J|i}(\theta_{J|i})$ are Dirichlet. Similarly, we have shown that $f_J(\theta_{\cdot J})$ and $f_{I|j}(\theta_{I|j})$ are Dirichlet. We now show that $f_U(\{\theta_{ij}\})$ is Dirichlet. This completes the proof of Theorem 2.

We can write

$$f_{JI}(\theta_{\cdot J}, \theta_{I|1}, \dots, \theta_{I|k}) = f_J(\theta_{\cdot J}) \prod_{j=1}^{n} f_{I|j}(\theta_{I|j}) = c \prod_{j=1}^{k} \theta_{\cdot j}^{\alpha_j - 1} \prod_{j=1}^{k} \prod_{i=1}^{n} \theta_{i|j}^{\alpha_{i|j} - 1}.$$

However, $f_{IJ}(\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k})$ can be expressed using f_{JI} by two applications of the Jacobian given by (3). Thus we get

$$\begin{aligned} &f_{IJ}(\theta_{I\cdot}, \theta_{J|1}, \dots, \theta_{J|k}) \\ &\quad = c\left[\prod_{i=1}^{k} \theta_{i\cdot}^{n-1} \right] \left[\prod_{j=1}^{n} \theta_{\cdot j}^{k-1} \right]^{-1} \left[\prod_{j=1}^{k} \theta_{\cdot j}^{\alpha_j - 1} \prod_{j=1}^{k} \prod_{i=1}^{n} \left[\frac{\theta_{j|i} \theta_{i\cdot}}{\theta_{\cdot j}} \right]^{\alpha_{i|j} - 1} \right], \end{aligned} \tag{80}$$

where $\theta_{\cdot j} = \sum_i \theta_{i\cdot} \theta_{j|i}$. Because f_{IJ} is a product of Dirichlet functions $f_I, f_{J|1}, \dots, f_{I|n}$, it follows from (80) that the exponent coefficients for $\theta_{\cdot j}$, $1 \le j \le n$, must vanish. Consequently, $f_U(\{\theta_{ij}\})$, which is obtained from (80) by multiplying with $\{\prod_{i=1}^{k} \theta_{i\cdot}^{n-1}\}^{-1}$ and using the relationship $\theta_{ij} = \theta_{j|i} \theta_{i\cdot}$, is Dirichlet.

Acknowledgments. We thank J. Aczél, M. Israeli and M. Ungarish for valuable comments. We thank S. Altschuler and L. Wu for their help with the proof of Lemma A.1 and A. Járai for referring us to the results in [13] and helping us understand their applicability to the problems discussed herein.

Note added in proof. Járai has recently shown that the assumption of strict positivity is redundant [Regularity property of the functional equation of the Dirichlet distribution. (1996). *Aequationes Mathematicae*. To appear.].

REFERENCES

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

[2] Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, *Los Angeles*, *July 1991* 52–60. Morgan Kaufmann, San Mateo, CA.

[3] Chickering, D. (1995). A transformational characterization of equivalent Bayesian-network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, *Montreal*, *July 1995*. Morgan Kaufmann, San Mateo, CA.

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

[5] Darroch, J. N. and Ratcliff, D. (1971). A characterization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **66** 641–643.

[6] Dawid, P. and Lauritzen, S. (1993). Hyper Markov laws in statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317.

[7] DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[8] Fabius, J. (1973). Two characterizations of the Dirichlet distribution. *Ann. Statist.* **1** 583–587.

[9] Geiger, D. and Heckerman, D. (1995). A characterization of the Dirichlet distribution with application to learning Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, *Montreal*, *August 1995* 196–207. Morgan Kaufmann, San Mateo, CA.

[10] Geiger, D., Verma, T. S. and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20** 507–534.

[11] Heckerman, D. E. (1991). *Probabilistic Similarity Networks*. MIT Press.

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

[13] Járai, A. (1986). On regular solutions of functional equations. *Aequationes Math.* **30** 21–54.

[14] James, I. R. and Mosimann, J. E. (1980). A new characterization of the Dirichlet distribution through neutrality. *Ann. Statist.* **8** 183–189.

[15] Jensen, F. V., Lauritzen, S. L. and Olesen, K. G. (1990). Bayesian updating in recursive graphical models by local computations. *Computational Statistical Quarterly* **4** 269–282.

[16] Kagan, A. M., Linnik, Y. V. and Rao, C. R. (1962). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

[17] Lauritzen, S. L. (1982). *Lectures on Contingency Tables*. Univ. Aalborg Press.

[18] Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

[19] Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *J. Amer. Statist. Assoc.* **62** 30–44.

[20] Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* **29** 241–288.

[21] Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence* **32** 245–257.

[22] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

[23] PEARL, J. (1993). Belief networks revisited. *Artificial Intelligence* **59** 49–56.

[24] PEARL, J. (1993). Comment: graphical models, causality, and intervention. *Statist*. *Sci*. **8** 266–269.

[25] PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

[26] RAMACHANDRAN, B. and LAU, K. (1991). *Functional Equations in Probability Theory*. Academic Press, New York.

[27] RAO, C. R. and SHANBHAG, D. N. (1994). *Choquet-Deny Type Functional Equations with Applications to Stochastic Models*. Wiley, New York.

[28] SHACHTER, R. D. (1986). Evaluating influence diagrams. *Oper*. *Res*. **34** 871–882.

[29] SHACHTER, R. D. (1988). Probabilistic inference and influence diagrams. *Oper*. *Res*. **36** 589–604.

[30] SPIEGELHALTER, D., DAWID, A., LAURITZEN, S. and COWELL, R. (1993). Bayesian analysis in expert systems. *Statist*. *Sci*. **8** 219–282.

[31] SPIEGELHALTER, D. and LAURITZEN, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20** 579–605.

[32] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1995). *Causation*, *Prediction*, *and Search*. Springer, New York.

[33] VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, *Boston* 220–227. Morgan Kaufmann, San Mateo, CA.

[34] WERMOUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70** 537–552.

[35] WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

[36] WILKS, S. (1962). *Mathematical Statistics*. Wiley, New York.

[37] ZABELL, S. (1982). W. E. Johnson's "sufficientness" postulate. *Ann*. *Statist*. **10** 1091–1099.

COMPUTER SCIENCE DEPARTMENT
TECHNION
HAIFA 32000
ISRAEL
E-MAIL: dang@cs.technion.ac.il

MICROSOFT RESEARCH
REDMOND, WASHINGTON 98052
E-MAIL: heckerma@microsoft.com