A Characterization of the Dirichlet Distribution through Global and Local Parameter Independence

Dan Geiger
dang@cs.technion.ac.il
(Main affiliation:
Computer Science Department
Technion, Haifa 32000, Israel
dang@cs.technion.ac.il)

David Heckerman
heckerma@microsoft.com

November 1994 (Revised March 1996)

Technical Report
MSR-TR-94-16

Microsoft Research
Advanced Technology Division
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Summary

We provide a new characterization of the Dirichlet distribution. Let θ_{ij}, $1 \leq i \leq k, 1 \leq j \leq n$, 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 $\{\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{\cdot J}, \theta_{I|1}, \ldots, \theta_{I|n}\}$ are mutually independent (where $\theta_{\cdot J}$ and $\theta_{I|j}$ are defined analogously), and assuming strictly positive pdfs, 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.

Abbreviated title. Characterization of Dirichlet Distributions

Key words and phrases. Bayesian network, Characterization, Dirichlet distribution, Functional equation, Graphical model, Hyper-Markov law.

AMS Subject Classifications. Primary 62E10, 60E05; Secondary 62A15, 62C10, 39B99.

1 Introduction

A statistical model that represents a large collection of discrete random variables imposes a 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 statistical literature (e.g., [La82, We83, LS88, JLO89, SL90, DL93, Pe95]), the artificial intelligence and computer science literature (e.g., [Pe86, Pe87, GVP90, Pe93a]), as well as in operations research (e.g. [Sh86, Sh87]) and philosophy (e.g.,[SGS95]). For an introduction of graphical models see the books [Pe88, Wh90] 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 alone. Each node i in a Bayesian network represents a random variable s_i and the joint distribution satisfies

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

where $i_1, \ldots, i_k$ 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 network is the well known Markov chain over $\{s_i | 1 \le i \le n\}$ which represents the distribution $P(s_1, \ldots, s_n) = \prod_i p(s_i | s_{i-1})$. Since 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 became the dominant model in artificial intelligence for representing knowledge needed for reasoning tasks that require the explicit representation of uncertainty.

In recent years, it has been realized that although Bayesian networks can be constructed manually directly from an expert, often using advanced computerized elicitation techniques [He91], it is advantageous to use data both to update the parameters as well as to (perhaps) select a different graphical model. The latter problem has been addressed by several researchers who have investigated Bayesian methods for model selection when the models are Bayesian networks [Bu91, CH92, HGC95, SDLC93]. 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.

The Bayesian approach for model selection can be described as follows. Suppose we have a set of discrete random variables $\{s_1, \ldots, s_n\} = U$, and a set of cases $\{C_1, \ldots, C_m\} = \mathrm{D}$ where each case is an instance of some or of all the variables in U. We sometimes refer to D as a database. Let (B_S, B_P) be a Bayesian network, that is, B_S is a network structure (a directed acyclic graph), each node i of B_s is associated with a random variable s_i and B_P is a set of conditional distributions, $p(s_i|s_{i_1}, \ldots, s_{i_k})$, $1 \leq i \leq n$, where $s_{i_1}, \ldots, s_{i_k}$ are the variables corresponding to the parents of node i in B_S. (For more details, consult [Pe88]). The conditional probability table, when known with certainty, is the set of parameters associated with node i. Let B_S^h stand for the hypothesis that cases are drawn from a Bayesian network having the structure B_S. Then a Bayesian measure of the goodness-of-fit of a Bayesian-network structure B_S is $p(B_S^h|\mathrm{D}, \xi)$ given by $p(B_S^h|\mathrm{D}, \xi) = c \cdot p(B_S^h|\xi) p(\mathrm{D}|B_S^h, \xi)$ where c is a normalizing factor and ξ is the current state of knowledge.

To compute $p(\mathrm{D}|B_S^h, \xi)$ in closed form several assumptions were made. First, the database D is a multinomial sample from some Bayesian network (B_S, B_P) and the prior probability of each structure is positive, i.e., $P(B_S^h|\xi) > 0$ for every B_S. Second, for each network structure the parameters associated with one node are independent of the parameters associated with other nodes (global independence [SL90]) and the parameters associated within a node given one instance of its parents are independent of the parameters of that node given other instances of its parent nodes (local independence [SL90]). Third, if a node has the same parents in two distinct networks structures, then the distribution of the parameters associated with this node are identical in both structures (parameter modularity [HGC95]). Forth, each case is complete, namely, an instance of all the variables represented by the network. Fifth, the distribution of the parameters associated with each node 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. A relaxation of the assumption of complete cases was carried out by previous works (e.g., [SDLC93]). 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 probability distributions of all the parameters are 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 actually 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 the joint probability $p(s,t)$ from a sample of pairs of values (s_i, t_j) of s and t. The standard Bayesian approach to this statistical inference problem is to associate with $p(s_i, t_j)$ a parameter θ_{ij} (often called the multinomial parameter), 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 closely-related variants to this approach which can be described as follows.

Let $\theta_{i\cdot} = \sum_{j=1}^n \theta_{ij}$ stand for the multinomial parameter associated with $p(s = s_i)$ and let $\theta_{j|i} = \theta_{ij} / \sum_j \theta_{ij}$ stand for the multinomial parameter associated 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 $\{\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent and that each has a prior pdf. Now 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 $\theta_{\cdot j} = \sum_{i=1}^k \theta_{ij}$, $\theta_{i|j} = \theta_{ij} / \sum_i \theta_{ij}$, $\theta_{\cdot J} = \{\theta_{\cdot j}\}_{j=1}^{n-1}$ and $\theta_{I|j} = \{\theta_{i|j}\}_{i=1}^{k-1}$. Now we assume that $\{\theta_{\cdot J}, \theta_{I|1}, \ldots, \theta_{I|n}\}$ are mutually independent and that each has a prior pdf and 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 $\{\theta_{ij}\}$ is a Dirichlet pdf usually for pragmatic reasons. When such a choice is made, it can be shown that $\{\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are indeed mutually independent and that each has a prior Dirichlet pdf. Similarly, $\{\theta_{\cdot J}, \theta_{I|1}, \ldots, \theta_{I|n}\}$ are mutually independent and each has a prior Dirichlet pdf.

The result proved in this article is that if these independence assertions are assumed to hold, and under the assumption of (strictly) positive pdfs, then a prior Dirichlet pdf for $\{\theta_{ij}\}$ is the only possible choice. We conjecture that the assumption of strictly positive pdfs can be dropped without affecting the conclusion but we have not carried out a proof of this claim. In Section 2, we discuss the 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-ways tables to n way tables and the implication to learning Bayesian networks. Further extensions are described in Section 4. An analogous result for a normal sampling situation which characterizes the Wishart distribution is outlined in [GH95].

2 Background and Technical Summary

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

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

where $\phi_l = 1 - \sum_{i=1}^{l-1} \phi_i$ and α_i are positive constants (See, e.g., [De70, Wi62]).

We use the following conventions. Suppose $\{\theta_{ij}\}$, $1 \leq i \leq k$, $1 \leq j \leq n$, 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 $\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}$ and f_{JI} be the joint pdf of $\theta_{\cdot J}, \theta_{I|1}, \ldots, \theta_{I|n}$.

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 $\theta_{kn} = 1 - \sum_A \theta_{ij}$, $A = \{(i,j) | 1 \leq i, j \leq n, i \neq k \text{ or } j \neq n\}$, 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. Since both $\{\theta_{i\cdot}\}_{i=1}^{k}$ and $\{\theta_{j|i}\}_{j=1}^{n}$ are defined in terms of $\{\theta_{ij}\}$ and since $\theta_{ij} = \theta_{i\cdot}\theta_{j|i}$, 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

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

[HGC95].

The following lemma provides a known property of the Dirichlet distribution. A slightly weaker version is stated in [DL93] (Lemma 7.2).

Lemma 1 *Let $\{\theta_{ij}\}$, $1 \leq i \leq k$, $1 \leq j \leq n$, 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, $1 \leq i \leq k$, and $\{\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent.*

Proof: Set $\theta_{ij} = \theta_{i\cdot}\theta_{j|i}$ in Eq. 2, multiply by J_{kn}, and regroup terms. $\Box$

The main claim of this article is that, under the assumption of a positive pdf for $\{\theta_{ij}\}$, the converse holds as well. More specifically, we prove the following theorem (Proof 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 positive pdf $f_U(\{\theta_{ij}\})$. If $\{\theta_{I\cdot}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{\cdot J}, \theta_{I|1}, \ldots, \theta_{I|n}\}$ 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 Equation 3. Since 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,

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

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

Methods for solving functional equations such as Eq. 4, that is, finding all functions that satisfy them under different regularity assumptions, are discussed in [Ac66]. We use the following technique. First, we argue that any positive solution to Eq. 4 must be differentiable in any order (Aczél, 66, Section 4.2.2, "Deduction of differentiability from integrability"). Then we take repeated derivatives of Eq. 4 and obtain a differential equation the solution of which after appropriate specialization is the general solution of Eq. 4 (Aczél, 66, 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 $f'(x+y) = f'(x)$. Since the latter equality holds for all y, it follows that $f'(t)$ is constant and thus $f(t)$ is linear in t [Ac66]. This functional equation is one of Cauchy's fundamental equations and it immediately establishes the memoryless property of the exponential distribution (e.g.,[MO67]). In Eq. 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, Eq. 4 can be written as follows:

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

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. A solution of this equation is given in the Appendix.

Járai (1986) has extensively investigated the following type of functional equations:

$$f(t) = h(t, y, f_0(y), f_1(g_1(t, y)), \ldots, f_n(g_n(t, y))) \tag{6}$$

where $f, f_0, \ldots, f_n$ are unknown functions, $h, g_1, \ldots, g_n$ 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 [Ja86]). Since Equation 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 (since 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 Equation 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 [Ja86]). Actually some stronger results of this sort are proved in [Ja86]. 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 and therefore all measurable solutions are linear.

In solving Equation 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 Equation 7. We assume that the solutions are strictly positive and measurable. Since the logarithm of a positive measurable function is measurable, we can now use Járai's theorems and obtain that all positive measurable solutions of Equation 4 have infinitely many derivatives.

Járai's theorems are very useful in Statistical applications since 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 [DR71]. 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 since 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 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 indeed do, the joint pdf of X and Y can be written in two distinct ways. Equating these representations, as in Equation 4, forms a functional equation. This functional equation is of the type dealt 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 [Fa73, JM80]. 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 [KLR62, RL91, RS94]. 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 [Za82] and references therein). This characterization is based, loosely speaking, on the assumption that for exchangeable sequences the expectation of the parameter of the i-th category depends only on the counts of the i-th category and the total count. Our assumptions on the other hand, in particular, global independence, were originally made so as to facilitate a prior-to-posterior analysis [DL93]. 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, e.g., all members are unimodal and thus Dirichlet mixtures are sometimes preferable. This point is raised again in the next section.

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 [MO67], the functional equation that defines a multivariate exponential distribution $F(X, Y)$ through an extended version of the memoryless property,

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

$x_1, x_2, y > 0$, 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 loosing an important term of the general solution. This situation occurs since a regularity condition of Járai is violated (the rank of the matrix of the first derivatives wrt y, $\frac{\partial g_i(t,y)}{\partial y}$, for each g_i in Equation 6 must equal the dimensionality of the domain of f_i; Here, the mapping $y \to (y, y)$ fails to meet this condition since the rank is one rather than two).

3 Implications For Learning

We now explain how our characterization applies to learning Bayesian networks. We concentrate on Bayesian networks for two discrete 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, the structure $s \to t$, and the structure $t \to s$. The first structure B_0 corresponds to the assertion that s and t are independent, whereas the second structure B_1 and the third one B_2 assert that s and t are dependent; B_1 represents the factorization $p(s, t) = p(s)p(t|s)$ and 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 variables U are *equivalent* if they encode the same set of joint probability distributions for U.

For example, network structures in which every pair of nodes is connected (complete network structures) are equivalent because they both encode all joint distributions for U. Another example is given in Figure 1. Characterizations of equivalent Bayesian networks for discrete variables are obtained in [Ch95, VP90].

We have said that, given a set of variables U and a network structure B, the hypothesis B^h stands for the assertion that the joint distribution of U factors according to B. By this definition of network-structure hypothesis, if follows immediately that if two network structures B_1 and B_2 are equivalent, then $B_1^h = B_2^h$. Following HGC95, we call this property *hypothesis equivalence*.

We note that some researchers give Bayesian-network structure a causal interpretation (e.g., [SGS95, Pe95]). For example, it is common to associate the network $s \to t$ with

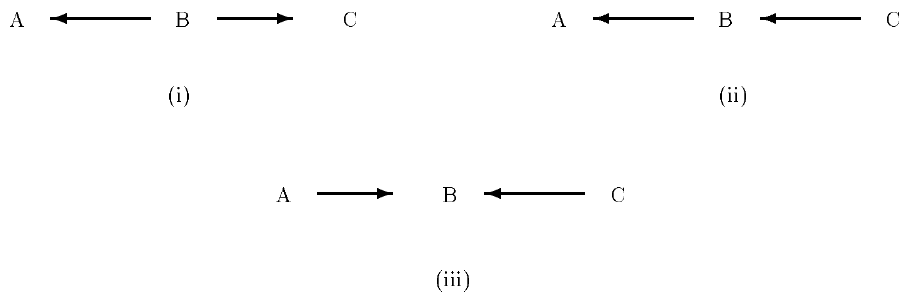

Figure 1: The pair of network structures (i) and (ii) are equivalent because both encode precisely those distributions where A and C are conditionally independent given B. Network structure (iii) is not isomorphic to them because it encodes those distributions where A and C are marginally independent.

the statement s causes t and the network $t \to s$ with the statement t causes s. Under this interpretation, the hypothesis B^h should correspond to the assertion that "the joint distribution of U factors according to B *and* the edges in B are in the causal direction." Given this alternative definition of network-structure hypothesis, the property of hypothesis equivalence no longer holds. That is, we can have $B_1^h \neq B_2^h$ even though B_1 and B_2 are equivalent.

Recalling the notations introduced in the introduction, we have that $\theta_{i\cdot} = \sum_{j=1}^{n} \theta_{ij}$ stand for the multinomial parameters associated with $p(s = s_i)$ and $\theta_{j|i} = \theta_{ij} / \sum_j \theta_{ij}$ stand for the multinomial parameters associated with $p(t = t_j | s = s_i)$. Given the first definition of network-structure hypothesis for which hypothesis equivalence holds, we obtain

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

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

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

Note that in this derivation we used the following assumptions in order to apply Theorem 2: a multinomial sampling situation, hypothesis equivalence, local and global independence, and the regularity conditions that each complete structure hypothesis has a positive probability (since we condition on B_i^h), and f_{IJ} is a strictly-positive pdf.

The parameter priors for the non-complete network structure B_0 follows from the assumption of parameter modularity, which says that if the nodes corresponding to a 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, modularity gives us

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

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

These equalities imply that the prior for the parameters of B_0 is Dirichlet as well.

Recall that the exponents of θ_{ij} of a Dirichlet distribution can be written as $N\alpha_{ij} - 1$ where N is the "equivalent sample size" (the size of an imaginary database of complete cases–the prior sample–upon which the prior Dirichlet is based) and α_{ij} is the expectation of θ_{ij}. The equivalent sample size reflects the confidence of the user and α_{ij} represents the relative frequency of the pair (i, j) in the prior sample. A joint Dirichlet prior is therefore quite restricting because it allows only one equivalent sample size for the entire domain. That is, there is no way to express different confidence levels regarding the parameters of different parts of the network. Thus the practical ramification of our characterization is that the commonly-made global and local 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 $\theta_{I\cdot}$ is more precise than knowledge about $\theta_{J|i}$.

One possibility for overcoming this limitation of the Dirichlet prior is to replace the notion of a single equivalent sample size with *equivalent database* [HGC95]. Namely, we ask a user to imagine that she was initially completely ignorant about a domain, having an uninformative prior with equivalent sample size(s) close to the lower bound. Then, we ask the user to specify a database D_e that would produce a posterior density that reflects her current state of knowledge. This database may contain incomplete cases. Then, to score a real database D, we score the database $\mathrm{D}_e \cup \mathrm{D}$, using the uninformative prior and a learning algorithm that handles missing data. This way of specifying a prior yields a mixture of Dirichlet distributions which, according to our result, cannot satisfy the local and global independence assumption.

This result of the inevitable choice of a Dirichlet prior for two-variables networks (two way tables) is easily generalized to the n-variate case (n way tables) 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 variables can be viewed as the sample space of a single discrete variable. Below we state the result in the notation of this article. For a proof consult [HGC95, Theorem 7].

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

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

For every configuration $i_1, \ldots, i_j$, $j = 1, \ldots, m$, we define, $\theta_{I_j|i_1,\ldots,i_{j-1}} = \{\theta_{i_j|i_1,\ldots,i_{j-1}}\}_{i_j=1}^{n_j-1}$ (for $j = 1$, $\theta_{I_j|i_1,\ldots,i_{j-1}} = \{\theta_{i_1}\}_{i_1=1}^{n_1-1}$). Similarly, we define

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

and let $\theta_{I_j|i_m,i_1,\ldots,i_{j-1}} = \{\theta_{i_j|i_m,i_1,\ldots,i_{j-1}}\}_{i_j=1}^{n_j-1}$ (for $j = m$, $\theta_{I_j|i_m,i_1,\ldots,i_{j-1}} = \{\theta_{i_m}\}_{i_m=1}^{n_m-1}$).

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

$$f(\{\theta_{i_1,\ldots,i_m}\}) \propto \prod_{i_1,\ldots,i_m} \theta_{i_1,\ldots,i_m}^{\alpha_{i_1,\ldots,i_m}-1} \tag{8}$$

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

$$\Theta_j = \{\theta_{I_j|i_1,\ldots,i_{j-1}} | 1 \leq i_1 \leq n_1, \ldots, 1 \leq i_{j-1} \leq n_{j-1}\}$$

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

$$\Phi_j = \{\theta_{I_j|i_m,i_1,\ldots,i_{j-1}} | 1 \leq i_m \leq n_m, 1 \leq i_1 \leq n_1, \ldots, 1 \leq i_{j-1} \leq n_{j-1}\}$$

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

4 Discussion

The independence assumptions made by Theorem 2 can be divided into two parts: $\{\theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{I|1}, \ldots, \theta_{I|n}\}$ are mutually independent (local independence) and $\theta_{I\cdot}$ is independent of $\{\theta_{J|1}, \ldots, \theta_{J|k}\}$ and $\theta_{\cdot J}$ is independent of $\{\theta_{I|1}, \ldots, \theta_{I|n}\}$ (global independence). A natural question to ask is whether global 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 [DL93]. Dawid and Lauritzen term a pdf that satisfies global 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 notations of Equation 5 the new functional equation can be written as follows:

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

where $x = yz + (1-y)w$. Note that Eq. 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 independence.

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

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

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 independence can be easily verified. It can in fact be shown, by solving Eq. 9, 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.

Since 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 so far proved this claim for two-variables 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 Equation 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 independence.

Acknowledgment

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 4 and A. Járai for referring us to the results in [Ja86] and helping us to understand their applicability to the problems discussed herein.

References

[Ac66] J. Aczél, *Lectures on functional equations and their applications*, 1966, Academic Press, New York.

[Bu91] W. Buntine, Theory refinement on Bayesian networks, Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence, Los Angeles, CA, July 1991, 52–60, Morgan Kaufmann.

[Ch95] D. Chickering, A transformational characterization of equivalent Bayesian-network structures, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal QU, July 1995. Morgan Kaufmann.

[CH92] G. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9:309–347, 1992.

[DR71] J.N. Darroch and D. Ratcliff, A characterization of the Dirichlet distribution, *Annals of Statistics*, 66:641–643, 1971.

[DL93] P. Dawid and S. Lauritzen, Hyper Markov laws in statistical analysis of decomposable graphical models, *Annals of Statistics*, 21:1272–1317, 1993.

[De70] M. DeGroot, *Optimal statistical decisions*, 1970, McGraw-Hill, New York.

[Fa73] J. Fabius, Two characterizations of the Dirichlet distribution, *Annals of Statistics*, 1:583-587, 1973.

[GH95] D. Geiger and D. Heckerman, A characterization of the Dirichlet distribution with application to learning Bayesian networks. Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QU, 196–207, August 1995. Morgan Kaufmann.

[GVP90] Geiger, D.; Verma, T.S.; and Pearl, J. Identifying independence in Bayesian networks. *Networks*, 20, 507–534, 1990.

[He91] D.E. Heckerman, *Probabilistic Similarity Networks*, 1991, MIT Press, Cambridge, MA.

[HGC95] D. Heckerman, D. Geiger and D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning, 20:197–243, 1995.

[JLO89] F.V. Jensen, S.L. Lauritzen and K.G. Olesen, Bayesian updating in recursive graphical models by local computations, *Computational Statistical Quarterly*, 4:269–282, 1990.

[Ja86] A. Járai, On regular solutions of functional equations, *Aequationes Mathematicae*, 30, 21–54, 1986.

[JM80] I. R. James and J. E. Mosimann, A new characterization of the Dirichlet distribution through neutrality, *Annals of Statistics*, 8:183-189, 1980.

[KLR62] A.M. Kagan, Y,V. Linnik and C.R. Rao, *Characterization problems in mathematical statistics*, 1962, John Wiley and Sons.

[La82] S.L. Lauritzen, *Lectures on Contingency Tables*, 1982, University of Aalborg Press, Aalborg, Denmark.

[LS88] S.L. Lauritzen and D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J. Royal Statistical Society B, 50:157–224, 1988.

[MO67] A. W. Marshall and I. Olkin, A multivariate exponential distribution, *Annals of Statistics*, 31–44, 1967.

[Pe86] J. Pearl, Fusion, propagation, and structuring in belief networks, *Artificial Intelligence*, 29:241–288, 1986.

[Pe87] J. Pearl, Evidential reasoning using stochastic simulation of causal models, *Artificial Intelligence*, 32:245–257, 1987.

[Pe88] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 1988, Morgan Kaufmann, San Mateo, CA.

[Pe93a] J. Pearl. Belief Networks Revisited, *Artificial Intelligence*, 59:49–56, 1993.

[Pe93b] J. Pearl. Comment: Graphical Models, Causality, and Intervention, *Statistical Science*, 8(2):266-269, 1993.

[Pe95] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669-710, 1995.

[Sh86] R.D., Shachter, Evaluating influence diagrams, *Operations Research*, 34:871-882, 1986.

[Sh87] R.D., Shachter, Probabilistic inference and influence diagrams, *Operations Research*, 36:589-604,1988.

[SGS95] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*, 1995, Springer-Verlag, New York.

[RL91] B. Ramachandran and K. Lau, *Functional equations in probability theory*, 1991, Academic press.

[RS94] C.R. Rao and D.N. Shanbhag, *Choquet-Deny type functional equations with applications to stochastic models*, 1994, John Wiley and Sons.

[SL90] D. Spiegelhalter, and S. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, *Networks*, 20, 579–605, 1990.

[SDLC93] D. Spiegelhalter, A. Dawid, S. Lauritzen and R. Cowell, Bayesian analysis in expert systems, *Statistical Science*, 8, 219–282, 1993.

[VP90] T. Verma, and J. Pearl, Equivalence and synthesis of causal models, Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence, Boston, MA, 220–227, 1990, Morgan Kaufmann.

[Wi62] S. Wilks, *Mathematical Statistics*, 1962, Wiley and Sons, New York.

[We83] N. Wermouth and S.L. Lauritzen, Graphical and recursive models for contingency tables, *Biometrika*, 70(3):537–552, 1983.

[Wh90] J. Whittaker, *Graphical models in applied multivariate statistics*, 1990, John Wiley and Sons.

[Za82] S. Zabell, W.E. Johnson's "sufficientness" postulate, *Annals of Statistics*, 10(4): 1091–1099, 1982.

Permanent address:

Dan Geiger, Computer Science Department, Technion, Haifa 32000, Israel.

David Heckerman, Microsoft Research, Redmond WA, 98052, USA.

Appendix: Proof

This appendix proves Theorem 2. The first section shows that we are allowed to take derivatives of the functional equation which Theorem 2 defines. The second section solves a special case of this functional equation, the third section gives some lemmas needed for the general solution, and the forth and fifth sections provide the general solution. The last section uses the general solution to complete the proof of Theorem 2.

The Functional Equation

By renaming of variable and function names, Eq. 4 can be written as follows:

$$f_0(y_1,\ldots,y_{n-1})\prod_{j=1}^{n} g_j(z_{1,j},\ldots,z_{k-1,j}) = g_0(x_1,\ldots,x_{k-1})\prod_{i=1}^{k} f_i(\frac{z_{i1}y_1}{x_i},\ldots,\frac{z_{i,n-1}y_{n-1}}{x_i}) \tag{11}$$

where

$$x_i = \sum_{j=1}^{n} z_{ij}y_j, \quad 1 \le i \le k-1 \tag{12}$$

$$z_{kj} = 1 - \sum_{i=1}^{k-1} z_{ij}, \qquad 1 \le j \le n$$

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 Eq. 11 are $y_1,\ldots,y_{n-1}$ (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 Eq. 11 are defined by Eqs. 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 $x_1\ldots,x_k$ (x_i replaces $\theta_{i\cdot}$) and $w_{ij} = \frac{z_{ij}y_j}{x_i}$, $1 \le i \le k$, $1 \le j \le n-1$ (w_{ij} replaces $\theta_{j|i}$) to be free variables and rewrite Eq. 11 in term of these variables, namely,

$$g_0(x_1,\ldots,x_{k-1})\prod_{i=1}^{k} f_i(w_{i,1},\ldots,w_{i,n-1}) = f_0(y_1,\ldots,y_{n-1})\prod_{j=1}^{n} g_j(\frac{w_{1,j}x_1}{y_j},\ldots,\frac{w_{k-1,j}x_{k-1}}{y_j}) \tag{14}$$

where

$$y_j = \sum_{i=1}^{k} w_{ij}x_i, \quad 1 \le j \le n-1 \tag{15}$$

$$w_{in} = 1 - \sum_{j=1}^{n-1} w_{ij}, \qquad 1 \le i \le k$$

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

We assume that all functions mentioned in Eq. 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 Eq. 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 Eq. 11, it is permissible to take any derivative at any point in the domain because it exists.

The Bivalued Equation

We shall now find all positive Lebesgue integrable solutions of Eq. 11 when $k = n = 2$. This derivation is different from the general derivation which is given in Section 4.

When $k = n = 2$, Eq. 11 reduces to:

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

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 Eq. 16 yields the following three equations,

$$\hat{f}_0'(y) - (z-w)\hat{g}_0'(x) = \frac{zw}{x^2}\hat{f}_1'(\frac{yz}{x}) + \frac{(1-z)(1-w)}{(1-x)^2}\hat{f}_2'(\frac{y(1-z)}{1-x}) \tag{20}$$

$$\hat{g}_1'(z) - y\hat{g}_0'(x) = \frac{yw(1-y)}{x^2}\hat{f}_1'(\frac{yz}{x}) - \frac{(1-w)(1-y)y}{(1-x)^2}\hat{f}_2'(\frac{y(1-z)}{1-x}) \tag{21}$$

$$\hat{g}_2'(w) - (1-y)\hat{g}_0'(x) = -\frac{yz(1-y)}{x^2}\hat{f}_1'(\frac{yz}{x}) + \frac{y(1-z)(1-y)}{(1-x)^2}\hat{f}_2'(\frac{y(1-z)}{1-x}) \tag{22}$$

Solving $\hat{f}_1'(\frac{yz}{x})$ and $\hat{f}_2'(\frac{y(1-z)}{1-x})$ from Eqs. 21 and 22, plugging the result back into Eq. 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,

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

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 Eq. 23 and multiplying the result by $1-y$, similarly, taking a derivative wrt w of Eq. 23 and multiplying the result by y yields, after subtracting the two equations,

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

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 Eq. 24 yields,

$$(1-2w)\hat{g}_2'(w) + w(1-w)\hat{g}_2''(w) = \frac{1-y}{y}\left[(1-2z)\hat{g}_1'(z) + z(1-z)\hat{g}_1''(z)\right] + (1-y)\hat{f}_0'(y) \tag{25}$$

Since 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. Eq. 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 $\frac{b}{w(1-w)}$ 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 $\alpha = b - \frac{c_1}{2}$ and $\beta = -(b + \frac{3c_1}{2})$. Hence, $g_2(w) = cw^{\alpha}(1-w)^{\beta}$ where c is a third arbitrary constant.

From Eq. 25 we also get,

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

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. In Section 4 we investigate the precise relationships between these constants.

Preliminary Lemmas

We now provide several lemmas that are needed for the derivation of the general solution of Eq. 11.

Lemma 4 *The general solution of the following partial differential equation for $f(x_1, \ldots, x_n)$,*

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

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

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

or, equivalently, by

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

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

Proof: Let $s = x_i$ and $t = \frac{x_i}{x_j}$. Thus, $f_{x_i} = f_s + \frac{t}{s} f_t$, $f_{x_j} = -\frac{t^2}{s} f_t$. Hence, after a change of variables, the differential equation becomes

$$f + s f_s = 0$$

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

Lemma 5 *The general solution of the following partial differential equation for* $f(x_1, \ldots, x_n)$*,*

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

is given by

$$f(x_1, \ldots, x_n) = \alpha \ln x_i - \beta \ln x_j + h(x_i + x_j, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \tag{31}$$

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

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

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

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

where g *is an arbitrary twice-differentiable function.*

Proof can be found in [GH95].

The General Solution

We now solve Eq. 11 for any n and k. First we assume both n and k are strictly greater than 2. We use the following notations:

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

Also we use the following notations:

$$\begin{aligned}
X &= (x_1, \ldots, x_{k-1}), & Z_j &= (z_{1,j}, \ldots, z_{k-1,j}), \\
Y &= (y_1, \ldots, y_{n-1}), & W_i &= (\tfrac{z_{i1}y_1}{x_i}, \ldots, \tfrac{z_{i,n-1}y_{n-1}}{x_i})
\end{aligned} \tag{34}$$

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

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

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

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

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

Taking now the logarithm and then a derivative wrt z_{in} $(1 \le i \le k-1)$ of Eq. 11 yields,

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

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

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

Subtracting Eq. 38 from Eq. 36 and setting $j = j_1$, yields,

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

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

Now we take a derivative wrt $z_{i_1 j_1}$ of Eq. 39 and obtain,

$$
\begin{aligned}
& \frac{1}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1 i_1} - g_{j_1}(Z_{j_1})_{i_2 i_1}\right] = \\
& -\frac{y_{j_1}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1} + \frac{y_{j_1}}{x_{i_1}} \sum_{l=1}^{n-1} f_{i_1}(W_{i_1})_{j_1 l}\left[-\frac{z_{i_1 l}y_l}{x_{i_1}^2}\right] + \frac{y_{j_1}}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1 j_1}
\end{aligned}
\tag{40}
$$

Similarly, we take a derivative wrt $z_{i_1 n}$ of Eq. 39 and obtain,

$$-\frac{1}{y_n}\left[g_n(Z_n)_{i_1 i_1} - g_n(Z_n)_{i_2 i_1}\right] = -\frac{y_n}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1} + \frac{y_n}{x_{i_1}} \sum_{l=1}^{n-1} f_{i_1}(W_{i_1})_{j_1 l}\left[-\frac{z_{i_1 l} y_l}{x_{i_1}^2}\right] \quad (41)$$

Eqs. 40 and 41 yield,

$$\frac{1}{y_{j_1}^2}\left[g_{j_1}(Z_{j_1})_{i_1 i_1} - g_{j_1}(Z_{j_1})_{i_2 i_1}\right] + \frac{1}{y_n^2}\left[g_n(Z_n)_{i_1 i_1} - g_n(Z_n)_{i_2 i_1}\right] = \frac{1}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1 j_1} \quad (42)$$

Now, we take a derivative wrt $z_{i_1 j_2}$ of Eq. 39 where $1 \leq j_2 \leq n-1$, $j_2 \neq j_1$ $(n \geq 3)$, and obtain,

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

Eqs. 41 and 43 yield $(j_1 \neq j_2)$,

$$\frac{1}{y_n^2}\left[g_n(Z_n)_{i_1 i_1} - g_n(Z_n)_{i_2 i_1}\right] = \frac{1}{x_{i_1}^2} f_{i_1}(W_{i_1})_{j_1 j_2} \quad (44)$$

Putting Eqs. 42 and 44 into Eq. 43 and recalling (from Eq. 12) that

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

we get,

$$\frac{1}{x_{i_1}} f_{i_1}(W_{i_1})_{j_1} = -\frac{z_{i_1 j_1}}{y_{j_1}}\left[g_{j_1}(Z_{j_1})_{i_1 i_1} - g_{j_1}(Z_{j_1})_{i_2 i_1}\right] + \frac{z_{i_1 n}}{y_n}\left[g_n(Z_n)_{i_1 i_1} - g_n(Z_n)_{i_2 i_1}\right] \quad (45)$$

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

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

Plugging Eqs. 45 and 46 into Eq. 39 and collecting all terms involving y_n in 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,

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

where $1 \leq j \leq n$.

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

$$\left[g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}\right] + z_{i_1 j}\left[g_j(Z_j)_{i_1 i_1} - g_j(Z_j)_{i_2 i_1}\right] + z_{i_2 j}\left[g_j(Z_j)_{i_1 i_2} - g_j(Z_j)_{i_2 i_2}\right] = 0 \quad (48)$$

Let $h(Z_j) = g_j(Z_j)_{i_1} - g_j(Z_j)_{i_2}$. Thus Eq. 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 4 provides the general solution for h and thus,

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

where

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

and where $\tilde{g}_j$ is an arbitrary function having one argument less than g_j, or also by Lemma 4,

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

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

$$f_i(W_i)_{j_1} - f_i(W_i)_{j_2} = \frac{x_i}{z_{ij_1} y_{j_1}}\tilde{f}_i(\frac{z_{ij_1} y_{j_1}}{z_{ij_2} y_{j_2}}, W_{j_1 j_2, i}) \tag{52}$$

where

$$W_{j_1 j_2, i} = (\frac{z_{i1} y_1}{x_i}, \ldots, \frac{z_{i, j_1 - 1} y_{j_1 - 1}}{x_i}, \frac{z_{i, j_1 + 1} y_{j_1 + 1}}{x_i}, \ldots, \frac{z_{i, j_2 - 1} y_{j_2 - 1}}{x_i}, \frac{z_{i, j_2 + 1} y_{j_2 + 1}}{x_i}, \ldots, \frac{z_{in} y_n}{x_i})$$

or also, we have,

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

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

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

Plugging Eqs. 50 through 53 into Eq. 54 yields,

$$\begin{aligned} &\frac{1}{z_{i_1 j_1} y_{j_1}}\tilde{g}_{j_1}(\frac{z_{i_1 j_1}}{z_{i_2 j_1}}, Z_{i_1 i_2, j_1}) - \frac{1}{z_{i_2 j_2} y_{j_2}}\tilde{g}_{j_2}(\frac{z_{i_1 j_2}}{z_{i_2 j_2}}, Z_{i_1 i_2, j_2}) = \\ &\quad \frac{1}{z_{i_1 j_1} y_{j_1}}\tilde{f}_{i_1}(\frac{z_{i_1 j_1} y_{j_1}}{z_{i_1 j_2} y_{j_2}}, W_{j_1 j_2, i_1}) - \frac{1}{z_{i_2 j_2} y_{j_2}}\tilde{f}_{i_2}(\frac{z_{i_2 j_1} y_{j_1}}{z_{i_2 j_2} y_{j_2}}, W_{j_1 j_2, i_2}) \end{aligned} \tag{55}$$

Note that the variables in $Z_{i_1 i_2, j_1}$ 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 Eq. 55 can be rewritten as follows,

$$\begin{aligned} \frac{1}{z_{i_1 j_1} y_{j_1}} \tilde{\tilde{g}}_{j_1}(\frac{z_{i_1 j_1}}{z_{i_2 j_1}}) - \frac{1}{z_{i_2 j_2} y_{j_2}} \tilde{\tilde{g}}_{j_2}(\frac{z_{i_1 j_2}}{z_{i_2 j_2}}) = \\ \frac{1}{z_{i_1 j_1} y_{j_1}} \tilde{\tilde{f}}_{i_1}(\frac{z_{i_1 j_1} y_{j_1}}{z_{i_1 j_2} y_{j_2}}) - \frac{1}{z_{i_2 j_2} y_{j_2}} \tilde{\tilde{f}}_{i_2}(\frac{z_{i_2 j_1} y_{j_1}}{z_{i_2 j_2} y_{j_2}}) \end{aligned} \tag{56}$$

Let, $x = z_{i_1 j_1} y_{j_1}$, $y = z_{i_2 j_1} y_{j_1}$, $z = z_{i_1 j_2} y_{j_2}$, and $w = z_{i_2 j_2} y_{j_2}$ in Eq. 56. Then,

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

By taking a derivative wrt y of Eq. 57, we get,

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

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

Eq. 59 is a first-order partial differential equation the general solution of which is given by Lemma 5. Consequently, due to Eq. 33, we get,

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

Now due to Lemma 6, we have,

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

Similarly,

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

which is obtained by repeating the derivation starting at Eq. 14 rather then at Eq. 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 Eq. 11 let $z_{1j} = z_{2j} = \cdots = z_{kj}$ for $1 \leq j \leq n$. Thus, according to Eq. 12, $z_{ij} = x_i$. Consequently, we get,

$$f_0(y_1, \ldots, y_{n-1}) \prod_{j=1}^{n} g_j(x_1, \ldots, x_{k-1}) = g_0(x_1, \ldots, x_{k-1}) \prod_{i=1}^{k} f_i(y_1, \ldots, y_{n-1}) \tag{63}$$

Eqs. 11 and 63 yield,

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

Plugging Eqs. 61 and 62 into Eq. 64, we get,

$$\left[\prod_{j=1}^{n}\prod_{i=1}^{k-1}(\frac{z_{ij}}{x_i})^{\alpha_{ij}}\right]\left[\prod_{j=1}^{n}\frac{G_j(\sum_{i=1}^{k-1} z_{ij})}{G_j(\sum_{i=1}^{k-1} x_i)}\right] = \left[\prod_{i=1}^{k}\prod_{j=1}^{n-1}(\frac{z_{ij}}{x_i})^{\beta_{ij}}\right]\left[\prod_{i=1}^{k}\frac{F_i(\sum_{j=1}^{n-1}\frac{z_{ij}y_j}{x_i})}{F_i(\sum_{j=1}^{n-1} y_j)}\right] \tag{65}$$

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

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

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

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

and where $\tilde{F}_k(t) = F_k(t)$ and $\tilde{G}_n(t) = G_n(t)$. We will show that $\tilde{F}_i(t)$, $i = 1, \ldots, k-1$, are constants. Consequently, due to Eq. 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 Eq. 12 and repeating the same arguments. By symmetric arguments each g_j also has a Dirichlet functional form.

Let $y_j = \frac{1}{n}$, for all j, $1 \leq j \leq n$ and $z_{ij} = \frac{1}{k}$ for all i and j, $1 \leq i \leq k$, $1 \leq j \leq n-1$. Hence, the only free variables remaining in Eq. 66 are z_{in} where $1 \leq i \leq k-1$. Note that $x_i = \sum_{j=1}^{n} z_{ij}y_j = \frac{n-1}{kn} + \frac{1}{n}z_{in}$, $1 \leq i \leq k-1$, 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 $1 \leq j \leq n-1$ and a function of $\sum_{i=1}^{k-1} z_{in}$ for $j = n$. Consequently, Eq. 66 becomes,

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

where $c = \frac{n-1}{kn}$, $d = \frac{1}{n}$ 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 k-th term on the right hand side of Eq. 66 is absorbed, along with some constants, into the definition of f in Eq. 67.

Let $t_i = \frac{c}{c+dz_{in}}$; $\quad z_{in} = \frac{c}{d}\frac{1-t_i}{t_i}$. Taking the logarithm of Eq. 67, we get,

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

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

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

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

$$\tilde{F}_i(t) = c_i t^{a_i} e^{\frac{K}{t}}, \quad 1 \leq i \leq k-1 \tag{70}$$

where K is a constant not depending on i.

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

$$g(\sum_{i=1}^{k-1} z_{i1}) = \prod_{i=1}^{k-1}\frac{(z_{i1}+w_0)^{a_i}}{z_{i1}^{c_{i1}}}e^{K\sum_{i=1}^{k-1}\frac{1}{z_{i1}+w_0}}$$

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

Thus,

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

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

Special Cases

We now solve Eq. 11 when $n=2$ and $k \geq 3$. This proof follows the general lines presented in the previous section but circumvents the applications of the fact $n \geq 3$ assumed in the previous section. When $k=3$ and $n \geq 3$, a similar derivation can be obtained, as implied by the symmetric roles of n and k in Eq. 4.

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

$$\frac{f_{i_1}(W_{i_1})_{j_1}}{x_{i_1}} = \frac{z_{i_1n}}{y_n}[g_n(Z_n)_{i_1i_1} - g_n(Z_n)_{i_2i_1}] - \frac{z_{i_1j_1}}{y_{j_1}}[g_{j_1}(Z_{j_1})_{i_1i_1} - g_{j_1}(Z_{j_1})_{i_2i_1}] \tag{73}$$

Similarly,

$$\frac{f_{i_2}(W_{i_2})_{j_1}}{x_{i_2}} = \frac{z_{i_2 n}}{y_n}[g_n(Z_n)_{i_1 i_2} - g_n(Z_n)_{i_2 i_2}] - \frac{z_{i_2 j_1}}{y_{j_1}}[g_{j_1}(Z_{j_1})_{i_1 i_2} - g_{j_1}(Z_{j_1})_{i_2 i_2}] \tag{74}$$

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

Plugging Eqs. 73 and 74 into Eq. 39 and collecting all terms involving y_n in 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 $g_j(Z_j)$, $1 \leq j \leq n$, given by Eq. 47. Consequently, as given by Eq. 50 and because $n = 2$,

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

and,

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

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

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

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

Plugging Eqs. 75 and 78 into Eq. 39 yields,

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

This equation parallels Eq. 56 where (j_2 is replaced by n) and can be solved in the same way. Thus Eq. 61 is obtained. Eq. 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 rest of the derivation follows closely the previous section.

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, $f_J(\theta_{\cdot J})$ and $f_{I|j}(\theta_{I|j})$ have been shown to be Dirichlet as well. We now show that $f_U(\{\theta_{ij}\})$ is Dirichlet. This completes the proof of Theorem 2.

We can write,

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

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

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

where $\theta_{\cdot j} = \sum_i \theta_{i\cdot}\theta_{j|i}$. Since f_{IJ} is a product of Dirichlet functions f_I, $f_{J|1}, \ldots, f_{I|n}$, it follows from Eq. 80 that the exponent coefficients for $\theta_{\cdot j}$, $1 \leq j \leq n$, must vanish. Consequently, $f_U(\{\theta_{ij}\})$, which is obtained from Eq. 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.