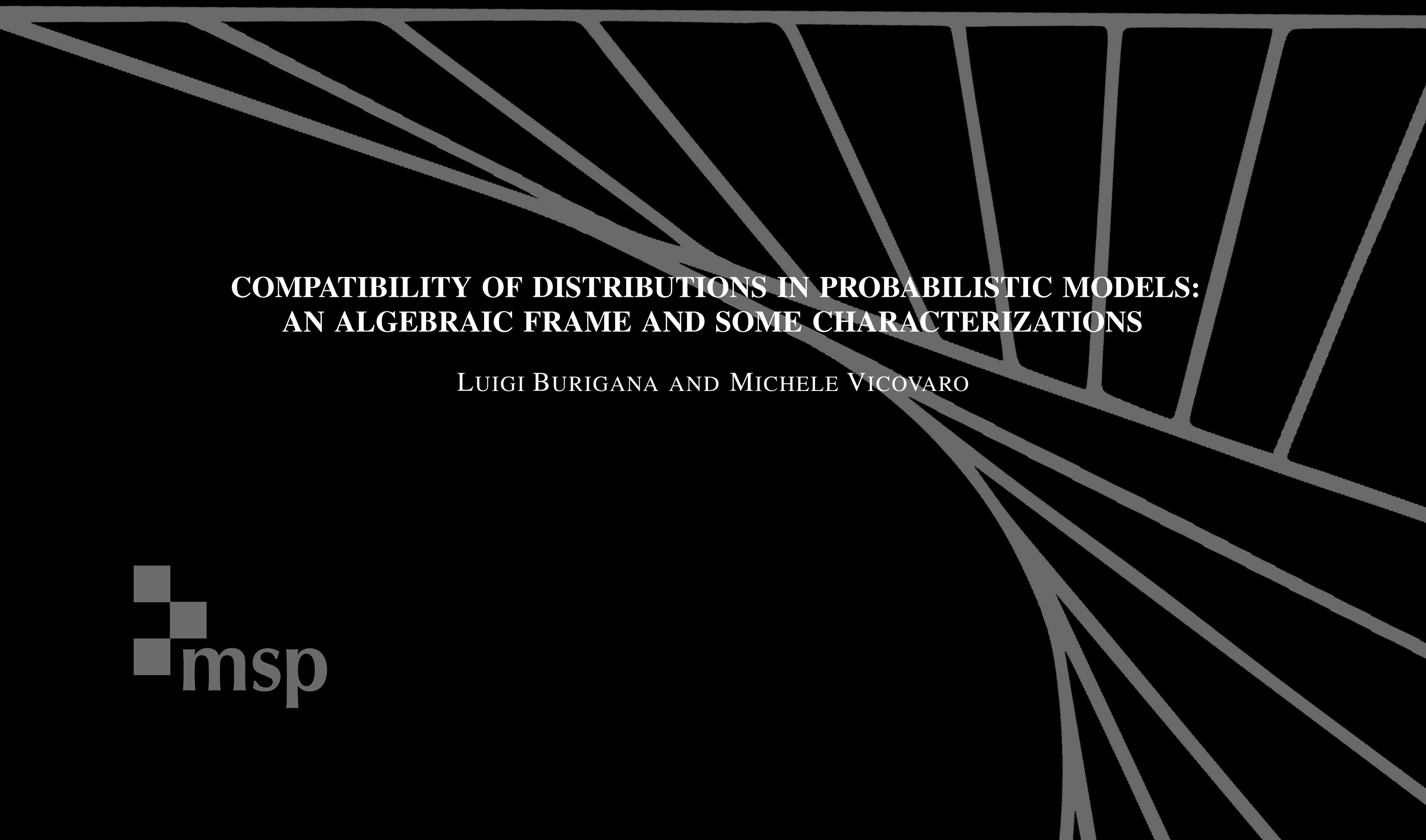
11 : 2 2020

Algebraic Statistics



COMPATIBILITY OF DISTRIBUTIONS IN PROBABILISTIC MODELS: AN ALGEBRAIC FRAME AND SOME CHARACTERIZATIONS

Luigi Burigana and Michele Vicovaro

msp

Algebraic Statistics
Vol. 11, No. 2, 2020
https://doi.org/10.2140/astat.2020.11.213

COMPATIBILITY OF DISTRIBUTIONS IN PROBABILISTIC MODELS: AN ALGEBRAIC FRAME AND SOME CHARACTERIZATIONS

LUIGI BURIGANA AND MICHELE VICOVARO

A probabilistic model may be formed of distinct distributional assumptions, and these may specify admissible distributions on distinct (not necessarily disjoint) subsets of the whole set of random variables of concern in the model. Such distributions on subsets of variables are said to be mutually compatible if there exists a distribution on the whole set of variables that precisely subsumes all of them. In Section 2 of this paper, an algebraic frame for this compatibility concept is constructed, by first observing that all marginal and/or conditional distributions (also called "probability kernels") that are implicit in a global distribution form a lattice, and then by highlighting the properties of useful operations that are internal to this algebraic structure. In Sections 3, 4, and 5, characterizations of the concept of compatibility are presented; first a characterization that depends only on set-theoretic relations between the variables involved in the distributions under judgment; then characterizations that are applicable only to pairs of candidate distributions; and then a characterization that is applicable to any set of candidate distributions when the variables involved in each of these are exhaustive of the set of variables in the model. Lastly, in Section 6, different categories of models are mentioned (a model of classical statistics, a corresponding hierarchical Bayesian model, Bayesian networks, Markov random fields, and the Gibbs sampler) to illustrate why the compatibility problem may have different levels of saliency and solutions in different kinds of probabilistic models.

1. Introduction

Several of the probabilistic models used in statistics and in other areas of applied probability are presented in modular form. By this, we mean that a model may be defined by a number of assumptions $A_1, \dots, A_m$ concerning the distributions acting on definite subsets $X_1, \dots, X_m$ of the total set $T = \{T_1, \dots, T_n\}$ of the elementary random variables of concern in the model. Some of these subsets may be disjoint, and others may overlap with one another. Any assumption A_i may specify a single distribution $p_i(X_i)$ or, more typically, it may fix only a constraint on an unknown $p_i(X_i)$ such that it in fact specifies a class of admissible distributions for X_i. Furthermore, each distribution $p_i(X_i)$ imposed or allowed for by an assumption A_i may be a marginal distribution; alternatively, it may be a conditional distribution that expresses how some of the variables in the set X_i are expected to be stochastically influenced by some other variables within the same X_i. Assumptions that specify or constrain the local conditional distributions are characteristic of hierarchical Bayesian models, Bayesian networks, Markov random fields, and probabilistic graphical models in general (Lauritzen, 1996; Koller and Friedman, 2009).

Burigana is the corresponding author.

MSC2020: 62H10, 62H22, 62R01, 60E99.

Keywords: probability kernel, conditional distribution, compatibility, lattice, graphical model.

When considering a model that is presented in modular form, the following question naturally arises. Suppose that $p_1(X_1), \dots, p_m(X_m)$ are local distributions specified (or allowed for) by the assumptions constituting the model. Are we assured that there exists a global distribution $p(T)$ that acts on the total set of variables and faithfully "assembles" these local distributions, in the sense that each $p_i(X_i)$ can be deduced from $p(T)$ through marginalization and/or conditioning? This is known as the *compatibility problem* for distributional assumptions (Berti, Dreassi and Rigo, 2014, p. 191). Compatibility is an essential requirement for the consistency and plausibility of a model as a whole. Indeed, if the local distributions $p_1(X_1), \dots, p_m(X_m)$ that comply with these assumptions were not mutually compatible, then analyses guided by the model (so far as these concern the whole set T of variables) would be disqualified as efforts towards a non-existing target, as there would be no $p(T)$ consistent with all $p_1(X_1), \dots, p_m(X_m)$. The compatibility problem thus conceived has been the subject of systematic research over the past three decades (Arnold, Castillo and Sarabia, 1999, 2001). Interest in this problem is particularly related to the study of so-called "conditionally specified statistical models", as the difficulty of testing compatibility greatly increases when the local distributions under judgment are in conditional form and act on sets of variables that overlap with one another.

With this study, we intend to contribute to the discussion of the compatibility requirement by setting this concept within an algebraic frame and presenting characterizations of it, some of which are taken and reformulated from the literature, and others we believe are new. The algebraic frame is defined in Section 2, and relies on the lattice structure possessed by the set of all marginal and conditional distributions that are deducible from a full joint distribution $p(T)$. The characterizations are presented in the next three sections; we examine a characterization that depends only on set-theoretic relations between the variables in the distributions under consideration (Section 3); characterizations limited to pairs of conditional distributions (Section 4); and a characterization that is applicable to any collection of conditional distributions such that the variables involved in each distribution exhaust the total set T (Section 5). Finally, in Section 6, we comment on simple examples to illustrate that in probabilistic models of different kinds, the compatibility problem may attain different saliency and may require different arguments for its solution.

The main reason for characterizing the frame of this study as an "algebraic" one is that our principal analyses will be conducted by working on structures that are lattices, as defined in abstract algebra, and by discussing relations and operations of algebraic character definable within those structures. In particular, this aspect will become apparent in Sections 2 and 3. Research on the compatibility problem, however, has also produced studies that can be categorized as "algebraic" for a complementary reason, that is, the mathematical tools used in them are of a kind familiar to contemporary algebraic statistics, such as analytical and computational tools from the theory of polynomials and algebraic geometry. Selected references to these studies will be presented in Sections 4 and 5 of our paper.

2. An algebraic view of variable pairs and probability kernels

In dealing with any probabilistic model, we assume that the elementary random quantities (individual data, parameters, hyper-parameters, etc.) involved in the model are exhaustively enumerated in a set

$T = \{T_1, \ldots, T_n\}$. The word *variable* is used here for any subset of T, including the whole T (the full variable), the empty set $\varnothing$ (the empty variable), any singleton $\{T_i\}$ (an elementary variable, also denoted by T_i), and arbitrary multiple variables (that is, sets of two or more elements of T). As they are understood as subsets of T, arbitrary variables may be compared or combined in set-theoretical manner, so that if X and Y are variables, then $X \cup Y$, $X \cap Y$, or $X \setminus Y$ are also variables.

For each elementary variable T_i, we assume that a measure space $(T_i^\circ, \mathcal{B}_i, \mu_i)$ is specified, in which T_i° is a standard set that includes all possible values of T_i (e.g., T_i° could be the real axis, or a definite subset of this), $\mathcal{B}_i$ is a sigma-field of subsets of T_i°, and μ_i is a reference measure on this sigma-field. Through multiplication, this construction associated with elementary variables is inherited by multiple variables. Specifically, a definite measure space $(X^\circ, \mathcal{B}_X, \mu_X)$ may be associated with any variable $X = \{T_{i_1}, \ldots, T_{i_k}\} \subseteq T$, in which $X^\circ = T_{i_1}^\circ \times \cdots \times T_{i_k}^\circ$ is the product of the spaces characteristic of the individual components (the space X° includes all possible values of X), $\mathcal{B}_X = \mathcal{B}_{i_1} \times \cdots \times \mathcal{B}_{i_k}$ is the product of the corresponding sigma-fields, and $\mu_X = \mu_{i_1} \times \cdots \times \mu_{i_k}$ is the product of the reference measures defined on these sigma-fields (Billingsley, 1995, § 18).

Discussions of conditional probability distributions imply reference to ordered pairs $(Y|X)$ in which X and Y are disjoint variables. Specifically, the term X (on the right of the bar) has the role of the conditioning variable and may be empty, whereas the term Y (on the left of the bar) has the role of the conditioned variable and is non-empty. We call $(Y|X)$ a *variable pair* and denote by $\mathcal{O}(T)$ the collection of such pairs. Simple combinatorics shows that if n is the cardinality of T, then $3^n - 2^n$ is the cardinality of $\mathcal{O}(T)$. For the purposes of our analysis, we make no substantial difference between any pairs $(\varnothing|X)$ and $(\varnothing|U)$ that have the empty variable on the left: both are symbols of *one* formal entity, called *null variable pair* and generally denoted by $\perp$. The symbol $\widetilde{\mathcal{O}}(T)$ stands for the set $\mathcal{O}(T) \cup \{\perp\}$.

For the utilities that will appear in the next paragraphs, the following criterion for comparing variable pairs is adopted.

Definition 1. A non-null variable pair $(V|U)$ is *dominated* by another non-null variable pair $(Y|X)$ (notation $(V|U) \preceq (Y|X)$) if the inclusions $V \cup U \subseteq Y \cup X$ and $U \supseteq X$ are both true. Furthermore, the null variable pair is dominated by any other variable pair (that is, $\perp \preceq (Y|X)$ for all $(Y|X) \in \mathcal{O}(T)$).

For example, assuming $T = \{T_1, \ldots, T_5\}$, if $(V|U) = (T_1, T_2|T_4, T_5)$, $(Y|X) = (T_1, T_2, T_4|T_5)$, and $(Z|W) = (T_1, T_2|T_5)$, then both $(V|U) \preceq (Y|X)$ and $(Z|W) \preceq (Y|X)$, but neither $(V|U) \preceq (Z|W)$ (condition $V \cup U \subseteq Z \cup W$ is violated) nor $(Z|W) \preceq (V|U)$ (condition $W \supseteq U$ is violated). Figure 1 allows us to present a useful characterization of the dominance defined above. This figure describes the crossing between two generic pairs $(V|U)$ and $(Y|X)$, by labeling the intersections and differences between the variables constituting each pair (for example, A stands for the difference $Y \setminus (V \cup U)$, E for the intersection $Y \cap U$, etc.). In these terms, it is readily seen that

$$(V|U) \preceq (Y|X) \text{ if and only if } B \cup C \cup D \cup G = \varnothing.$$

Figure 1 also plays a crucial role in the proofs of Theorems 1 and 2 stated below.

From the stated definition, any relation $(V|U) \preceq (Y|X)$ is the logical conjunction of the inclusions $V \cup U \subseteq Y \cup X$ and $U \supseteq X$. Due to this and to the fact that inclusion is a partial order (a reflexive,

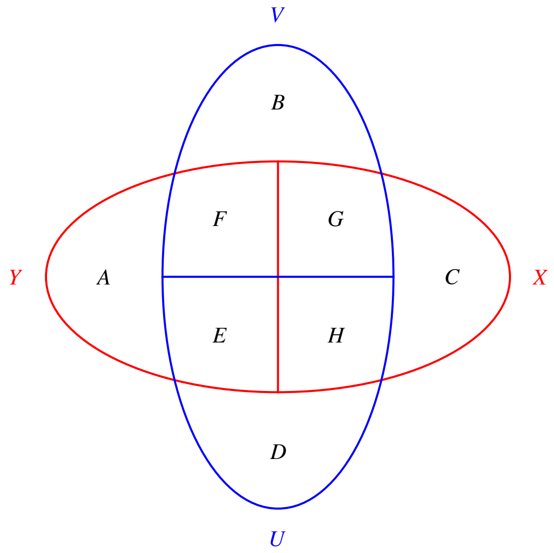


Figure 1. Crossing of two generic variable pairs $(V|U) = (B \cup F \cup G|D \cup E \cup H)$ and $(Y|X) = (A \cup E \cup F|C \cup G \cup H)$. Some of the eight subvariables $A, \ldots, H$ may be empty.

transitive, and antisymmetric relation), we obtain that the relation $\preceq$ is itself a partial order, and that the structure $(\widetilde{\mathcal{O}}(T), \preceq)$ is a partially ordered set. More specifically, the following properties can be proved (Burigana and Vicovaro, 2020, Proposition 1).

Proposition 1. *The structure* $(\widetilde{\mathcal{O}}(T), \preceq)$ *is a lattice in which the pair* $(T|\varnothing)$ *is the supremum, the null pair* $\bot$ *is the infimum, and for all non-null pairs* $(V|U)$ *and* $(Y|X)$ *their join* (*least upper bound*) *and meet* (*greatest lower bound*) *are given by the following equations*:

$$(V|U) \vee (Y|X) = (V \cup Y \cup (U + X)|U \cap X) \tag{1}$$
$$\text{with } U + X = (U \setminus X) \cup (X \setminus U);$$

$$(V|U) \wedge (Y|X) = (V \cap Y|U \cup X) \text{ or } = \bot \tag{2}$$
$$\text{depending on whether the conditions } V \cap Y \neq \varnothing,\ U \subseteq Y \cup X,\ X \subseteq V \cup U$$
$$\text{are or are not jointly true.}$$

The following additional properties are easily recognized: the atoms in the lattice $\widetilde{\mathcal{O}}(T)$ are the pairs $(Y|X)$ with $|Y| = 1$ (i.e., the conditioned variable is elementary, so that if $n = |T|$, then $n2^{n-1}$ is the number of atoms); the lattice is atomic, in that each member of $\mathcal{O}(T)$ can be expressed as the join of a suitable set of atoms; it is rankable, the rank of any pair $(Y|X)$ being simply the cardinality $|Y|$ of the conditioned variable; and it is locally distributive, by which we mean that the distributive laws hold true on any triple of variable pairs whose pairwise meets are all different from $\bot$ (the null variable pair). Figure 2 shows the Hasse diagrams (in three-dimensional form) of the lattices $\widetilde{\mathcal{O}}(T)$ for $|T| = 2, 3, 4$. In each diagram, a downward line represents a covering $(V|U) \prec\!\cdot (Y|X)$ in which $|Y \setminus V| = 1$ and

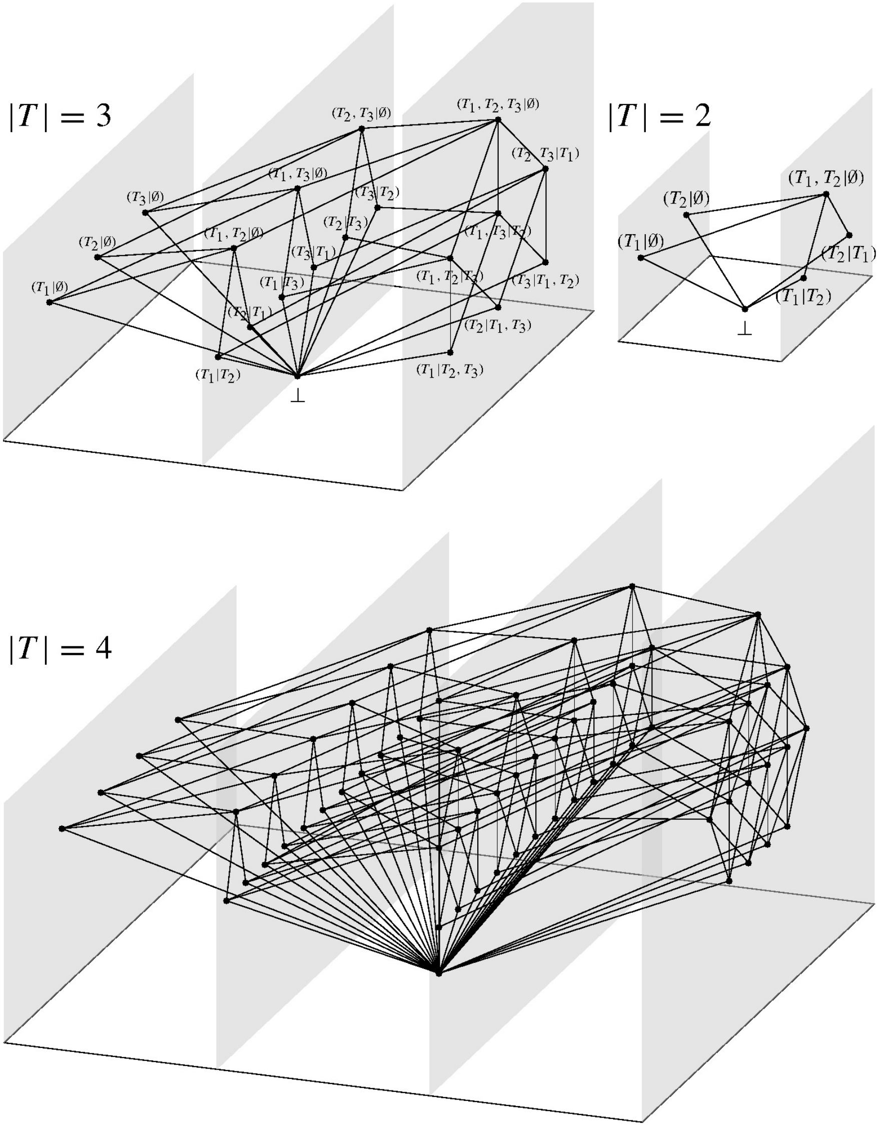


Figure 2. Three-dimensional Hasse diagrams of the lattices $\widetilde{O}(T)$ for $|T| = 2$ (top right), $|T| = 3$ (top left), and (bottom) $|T| = 4$.

$Y \setminus V = U \setminus X$ (that is, $(V|U)$ is derived from $(Y|X)$ by *transferring* one elementary variable from the left to the right component of the pair), whereas a backward line represents a covering $(V|U) \prec\!\cdot (Y|X)$ in which $|Y \setminus V| = 1$ and $U = X$ (that is, $(V|U)$ is derived from $(Y|X)$ by *cancelling* one elementary variable from the left component of the pair).

The next definition introduces the basic probabilistic objects of our study.

Definition 2. For any variable pair $(Y|X)$, a *probability kernel* associated with it is any family $\{p(Y|x) : x \in X^\circ\}$ (indexed by the values of X) in which each member $p(Y|x)$ is a non-negative valued function defined on the space Y°, measurable relative to the sigma-field $\mathcal{B}_Y$, and such that $\int p(y|x)\mu_Y(dy) = 1$. The family is also denoted by the symbol $p(Y|X)$.

The terms Y°, $\mathcal{B}_Y$, and μ_Y in this definition are the components of the measure space associated with the conditioned variable Y in the kernel. Following common usage, any member $p(Y|x)$ of a given family $p(Y|X)$ is referred to here as a *density*, irrespective of the kind (discrete, continuous, mixed, or other) of the variable Y[1]. The symbol $S(p(Y|x))$ will denote the *support* for any density $p(Y|x)$, that is, the set of points in the space Y° on which the density is positive. Furthermore, simply by considering the set-theoretic characteristics of the variables involved, basic kinds of kernels may be distinguished, which are assigned distinctive names. In particular, a kernel $p(Y|X)$ is *elementary* if Y is an elementary variable (a singleton in T), *saturated* if $X \cup Y = T$ (the full variable), *marginal* if $X = \varnothing$ (a marginal kernel $p(Y|\varnothing)$ is tantamount to one density $p(Y)$ on the space Y°), *full* if $Y = T$ (one density $p(T)$ over the space T°), *null* if $Y = \varnothing$ (symbol $\sharp$ is used here for the null kernel).

Probability kernels are subject to peculiar operations. The following are two basic exemplars.

Definition 3. Let $p(Y \cup Z|X)$ be a kernel in which Y and Z are non-empty disjoint variables.
(i) The result of *projecting* $p(Y \cup Z|X)$ relative to Z, denoted by $J[p(Y \cup Z|X), Z]$, is the kernel that is formed on the variable pair $(Y|X)$ by setting for all $(y, x) \in Y^\circ \times X^\circ$

$$p(y|x) = \int p(y, z|x)\mu_Z(dz).$$

(ii) The result of *conditioning* $p(Y \cup Z|X)$ relative to Z, denoted by $C[p(Y \cup Z|X), Z]$, is the kernel that is formed on the variable pair $(Y|Z \cup X)$ by setting for all $(y, z, x) \in Y^\circ \times Z^\circ \times X^\circ$

$$p(y|z, x) = \begin{cases} \dfrac{p(y, z|x)}{p(z|x)} & \text{if } p(z|x) \neq 0, \\ q(y) & \text{if } p(z|x) = 0, \end{cases}$$

where $p(Z|X) = J[p(Y \cup Z|X), Y]$ and $q(Y)$ is a freely chosen density on the space Y°.

The projection operation (symbol J) is tantamount to marginalization, as applicable to multivariate density functions. The conditioning operation (symbol C) is determined here as a division (the ratio $p(y, z|x)/p(z|x)$ in the formula), thus imitating the concept of conditional probability in its elementary version. For completeness, the definition may include an arbitrary density $q(Y)$, which however does not affect the univocal recovery of $p(Y \cup Z|X)$ from $p(Y|Z \cup X)$ and $p(Z|X)$ through the promotion operation in Definition 5 hereafter[2]. In the stated form, projection and conditioning are only defined for

[1] Based on any kernel $p(Y|X) = \{p(Y|x) : x \in X^\circ\}$, which according to Definition 2 is a family of point functions, a family $P_{Y|X} = \{P_{Y|x} : x \in X^\circ\}$ of set functions on the sigma-field $\mathcal{B}_Y$ can be constructed by setting $P_{Y|x}(B) = \int_B p(y|x)\mu_Y(dy)$ for all $B \in \mathcal{B}_Y$ and $x \in X^\circ$. In fact, in the measure-theoretic approach to probability, it is precisely a family like $P_{Y|X}$ that is called a probability kernel and is taken as a primitive structure, whereas $p(Y|X)$ is deduced as a corresponding family of Radon-Nikodym derivatives (Pollard, 2002, pp. 84, 119). Besides the term "probability kernel", other expressions are also used in different contexts to indicate probabilistic structures of the stated kind, such as "transition probability measure" (Parthasarathy, 2005, p. 174), "probability potential" (Koski and Noble, 2009, p. 58), "characteristic" (Griffeath, 1976, p. 426), "conditional probability distribution" (Koller and Friedman, 2009, p. 47), or simply "conditional" (Gelfand and Smith, 1990, p. 400).

[2] A discussion of the formal difficulties implicit in the intuitive notion of conditional probability and alternative ways of addressing them is given by Chang and Pollard (1997).

any $p(Y \cup Z|X)$ such that Y and Z are non-empty. Both concepts, however, may consistently be extended beyond this boundary by assuming these equations:

$$\begin{aligned} J[p(Y|X), \varnothing] &= p(Y|X) = C[p(Y|X), \varnothing], \\ J[p(Z|X), Z] &= \sharp = C[p(Z|X), Z]. \end{aligned} \tag{3}$$

For example, $J[p(Z|X), Z] = J[p(\varnothing \cup Z|X), Z] = p(\varnothing|X)$, which is the null kernel $\sharp$. Lastly, the following equations, in which X, Y, W, and Z are disjoint variables, are easily deduced from Definition 3:

$$\begin{aligned} J[J[p(Y \cup Z \cup W|X), Z], W] &= p(Y|X) = J[J[p(Y \cup Z \cup W|X), W], Z], \\ C[C[p(Y \cup Z \cup W|X), Z], W] &= p(Y|Z \cup W \cup X) = C[C[p(Y \cup Z \cup W|X), W], Z], \\ J[C[p(Y \cup Z \cup W|X), Z], W] &= p(Y|Z \cup X) = C[J[p(Y \cup Z \cup W|X), W], Z]. \end{aligned}$$

They express commutativity properties of projection and conditioning.

These operations determine a key relation among kernels.

Definition 4. A kernel $p(V|U)$ is *dominated* by a kernel $p(Y|X)$ (notation $p(V|U) \preceq p(Y|X)$) if the former can be obtained from the latter by projection, conditioning, or a combination of projection and conditioning.

Note that, if $p(V|U) \preceq p(Y|X)$, meaning that $p(V|U) = J[C[p(Y|X), W], Z]$ for some W and Z disjoint sub-variables of Y, then $V = Y \setminus (W \cup Z)$ and $U = W \cup X$ according to Definition 3, so that $V \cup U \subseteq Y \cup X$ and $U \supseteq X$, and therefore $(V|U) \preceq (Y|X)$ on applying Definition 1. In other words:

$$\text{if } p(V|U) \preceq p(Y|X) \text{ then } (V|U) \preceq (Y|X). \tag{4}$$

Hence, dominance between variable pairs (the symbol $\preceq$ in the consequent of this implication) is a necessary condition for dominance between kernels (the symbol $\preceq$ in the antecedent).

From a general perspective, given a full density $p(T) = p(T|\varnothing)$, we may consider the set of *all* kernels that are *dominated* by $p(T)$, a set denoted here by $\mathcal{P}(T)$. On the one hand, implication (4) shows that there is a natural one-to-one correspondence between this set $\mathcal{P}(T)$ and the set $\mathcal{O}(T)$ of all variable pairs in T, and there is also correspondence between the null kernel $\sharp$ and the null variable pair $\perp$. On the other hand, it can be seen that if comparison is limited to kernels belonging to a definite set $\mathcal{P}(T)$, then the implication (4) is reinforced as a bi-implication, that is:

$$\begin{aligned} &\text{for all } p(V|U) \text{ and } p(Y|X) \text{ in } \mathcal{P}(T) \\ &p(V|U) \preceq p(Y|X) \text{ if and only if } (V|U) \preceq (Y|X), \end{aligned}$$

so that the abovementioned correspondence is in fact an isomorphism between the structure $(\widetilde{\mathcal{P}}(T), \preceq)$ (with $\widetilde{\mathcal{P}}(T) = \mathcal{P}(T) \cup \{\sharp\}$) and the structure $(\widetilde{\mathcal{O}}(T), \preceq)$ (with $\widetilde{\mathcal{O}}(T) = \mathcal{O}(T) \cup \{\perp\}$). Proposition 1 ensures that the latter structure is a lattice. This means that the former is also a lattice, referred to here as the *lattice of kernels* generated by the assumed full density $p(T)$. This full density is the supremum in the lattice $\widetilde{\mathcal{P}}(T)$, while the infimum is the null kernel $\sharp$, and the atoms are the elementary kernels. The join and meet operations are expressed by formulas that are similar to (1) and (2) but involve arbitrary kernels $p(V|U)$ and $p(Y|X)$, rather than variable pairs. In the Hasse diagram of a lattice $\widetilde{\mathcal{P}}(T)$ (see Figure 2) each backward line represents the move from a $p(Y|X)$ to $p(Y \setminus \{T_i\}|X) = J[p(Y|X), \{T_i\}]$ by projection relative to an elementary variable $T_i \in Y$, whereas each downward line represents the move from a $p(Y|X)$ to $p(Y \setminus \{T_i\}|\{T_i\} \cup X) = C[p(Y|X), \{T_i\}]$ by elementary conditioning.

For later use, we note the following special property of any lattice of kernels.

Lemma 1. *Let $p(T)$ be a full density and suppose that $q(Y|Z \cup X)$ and $q(Y|X)$ are kernels such that $q(Y|z, x) = q(Y|x)$ for all $(z, x) \in Z^\circ \times X^\circ$. In these conditions, if $q(Y|Z \cup X)$ belongs to the lattice $\widetilde{\mathcal{P}}(T)$ generated by $p(T)$, then $q(Y|X)$ also belongs to the same lattice.*

Proof. Consider the kernels $p(Y|Z \cup X) = J[C[p(T), Z \cup X], T \setminus (Y \cup Z \cup X)]$ and $p(Y|X) = J[C[p(T), X], T \setminus (Y \cup X)]$, which surely belong to $\widetilde{\mathcal{P}}(T)$. If $q(Y|Z \cup X) \in \widetilde{\mathcal{P}}(T)$, then $q(Y|Z \cup X) = p(Y|Z \cup X)$, due to the one-to-one correspondence between $\widetilde{\mathcal{P}}(T)$ and $\widetilde{\mathcal{O}}(T)$. The hypothesized relation between $q(Y|Z \cup X)$ and $q(Y|X)$ implies that for each $x \in X^\circ$, the density $p(Y|z, x)$ is invariant relative to z varying in Z°. Thus, under the density $p(T)$ the variables Y and Z are conditionally independent given X, which means $p(Y|z, x) = p(Y|x)$ for all $(z, x) \in Z^\circ \times X^\circ$ (Lauritzen, 1996, p. 29). We then have $q(Y|x) = q(Y|z, x) = p(Y|z, x) = p(Y|x)$ for all $(z, x) \in Z^\circ \times X^\circ$, so that $q(Y|X) = p(Y|X)$, which combined with $p(Y|X) \in \widetilde{\mathcal{P}}(T)$ implies $q(Y|X) \in \widetilde{\mathcal{P}}(T)$. □

In our analyses, in addition to the projection and conditioning operations (which have a kernel and a variable as operands), two further operations are considered that have two kernels as operands. These are constrained operations, since to be applied they require that the operands (and, in particular, the variable pairs in these) comply with definite conditions.

Definition 5. (i) Given two kernels $p(Y|V \cup U)$ and $p(V|U)$, the result of *promoting* the former by the latter, denoted by $M[p(Y|V \cup U), p(V|U)]$, is the kernel that is formed on the variable pair $(Y \cup V|U)$ by setting for all $(y, v, u) \in Y^\circ \times V^\circ \times U^\circ$

$$p(y, v|u) = p(y|v, u) \cdot p(v|u).$$

(ii) Given two kernels $p(Y|V \cup U)$ and $p(V|Y \cup U)$ that are dominated by some full density $p(T)$ under which the equation $S(p(Y \cup V \cup U)) = S(p(Y)) \times S(p(V)) \times S(p(U))$ concerning the supports is satisfied, the result of *lightening* the former kernel by the latter, denoted by $L[p(Y|V \cup U), p(V|Y \cup U)]$, is the kernel that is formed on the variable pair $(Y|U)$ by setting for all $(y, u) \in Y^\circ \times U^\circ$

$$p(y|u) = \frac{1}{\displaystyle\int \frac{p(v|y, u)}{p(y|v, u)} \mu_V(dv)}.$$

With regard to the variable pairs, the applicability conditions of these two operations are implicit in the notation used in their definition. In particular, promotion is applicable only if the variable pair in the second operand $p(V|U)$ (the promoter) is a bipartition of the conditioning variable in the first operand $p(Y|V \cup U)$; in relation to this, the operation has the effect of *transferring* the variable V from the right to the left of the bar, yielding $p(Y \cup V|U)$ as the result. Lightening is applicable only if the variable pairs in the operands $p(Y|V \cup U)$ and $p(V|Y \cup U)$ have the same union and the conditioned variables are disjoint; the operation has the effect of *cancelling* the variable V from $p(Y|V \cup U)$, thus yielding $p(Y|U)$ as the result. Furthermore, lightening is applicable only if the factorability $S(p(Y \cup V \cup U)) = S(p(Y)) \times S(p(V)) \times S(p(U))$ of the support of $p(Y \cup V \cup U)$ is satisfied, meaning that for each $(y, v, u) \in Y^\circ \times V^\circ \times U^\circ$ the value $p(y, v, u)$ is positive if (and only if) all three values $p(y)$, $p(v)$, and $p(u)$ are positive, where $p(Y \cup V \cup U)$, $p(Y)$, $p(V)$, and $p(U)$ are densities deducible from some full density $p(T)$ through projection[3]. Under this condition, for each $(y, v, u) \in S(p(Y \cup V \cup U))$ the ratio $p(v|y, u)/p(y|v, u)$ does exist as a positive real number and for each $(y, u) \in S(p(Y \cup U))$ the following equations are true:

$$\frac{1}{\int \frac{p(v|y,u)}{p(y|v,u)} \mu_V(dv)} = \frac{1}{\int \frac{p(y,v,u)/p(y,u)}{p(y,v,u)/p(v,u)} \mu_V(dv)} = \frac{p(y,u)}{\int p(v,u)\mu_V(dv)} = \frac{p(y,u)}{p(u)} = p(y|u).$$

This shows that the result $L[p(Y|V \cup U), p(V|Y \cup U)]$ of lightening is precisely the kernel $p(Y|U)$ belonging to the same lattice to which the operands $p(Y|V \cup U)$ and $p(V|Y \cup U)$ belong[4].

With regard to the promotion operation, it is readily seen that for any two kernels $p(Y|V \cup U)$ and $p(V|U)$ its result $p(Y \cup V|U) = M[p(Y|V \cup U), p(V|U)]$ is itself a kernel on the indicated variable pair. In particular, for all $u \in U^\circ$,

$$\begin{aligned}\int p(y, v|u)(\mu_Y \times \mu_V)(d(y, v)) &= \int p(y|v, u) \cdot p(v|u)(\mu_Y \times \mu_V)(d(y, v)) \\ &= \int \Big[\int p(y|v, u)\mu_Y(dy) \Big] \cdot p(v|u)\mu_V(dv) = \int 1 \cdot p(v|u)\mu_V(dv) = 1\end{aligned}$$

where the second step is justified by Fubini's theorem. When reference is made to a definite lattice of kernels, the promotion operation (as well as the lightening operation) may then be viewed as a binary operation internal to the lattice and subject to a specific applicability condition. Indeed, when applicable, promotion produces the same results as the join operation in the lattice, as may be inferred from Equation (1). Furthermore, its definition may be consistently refined by assuming these equations:

$$M[\sharp, p(V|U)] = p(V|U), \qquad M[p(Y|X), \sharp] = p(Y|X). \tag{5}$$

These characterize the null kernel $\sharp$ as the left and right identity term for promotion and can be justified by replacing $\sharp$ by $p(\varnothing|V \cup U)$ in one case and by $p(\varnothing|X)$ in the other. Lastly, the following equation is easily deduced from Definition 5(i):

$$\begin{aligned}&M[M[p(Y|W \cup V \cup U), p(W|V \cup U)], p(V|U)] = p(Y \cup W \cup V|U) = \\ &M[p(Y|W \cup V \cup U), M[p(W|V \cup U), p(V|U)]].\end{aligned} \tag{6}$$

This characterizes promotion as an associative operation.

[3] This factorability requirement is also known as the "positivity condition" regarding multivariate densities (Besag, 1974, p. 195).

[4] The operation we call "lightening" was considered, for example, by Gourieroux and Monfort (1979) and Robert and Casella (2004, p. 344).

On the whole, we have four basic operations on probability kernels, with the symbols J (proJection), C (Conditioning), M (proMotion), and L (Lightening). In the next lemma, several equations are highlighted that arise from the combined use of these operations and will be applied in the following.

Lemma 2. *In each of the following equations, the kernels involved are assumed to belong to one lattice* $\widetilde{\mathcal{P}}(T)$.

(i) $J[M[p(Y|V \cup U), p(V|U)], V] = p(Y|U)$ (*relative to the first operand, part V of the conditioning variable is cancelled*).

(ii) $C[M[p(Y|V \cup U), p(V|U)], V] = p(Y|V \cup U)$ (*recovery of the first operand of a promotion*).

(iii) $J[M[p(Y|V \cup U), p(V|U)], Y] = p(V|U)$ (*recovery of the second operand of a promotion*).

(iv) $J[M[p(Y \cup X|V \cup U), p(V|U)], X] = M[J[p(Y \cup X|V \cup U), X], p(V|U)]$ (*a kind of commutativity between projection and promotion*).

(v) $M[p(Y|V \cup U), L[p(V|Y \cup U), p(Y|V \cup U)]] = p(Y \cup V|U)$ (*simulation of the join operation*).

(vi) $M[C[p(Y \cup Z|X), Z], J[p(Y \cup Z|X), Y)]] = p(Y \cup Z|X)$ (*recovery of a kernel through promotion*).

Proof. Each of these equations can be proved by noting that the kernel resulting from the composite operation on the left-hand side acts on the same variable pair as the kernel specified on the right-hand side, and then considering the one-to-one correspondence between variable pairs and kernels in a lattice (as implied by Equation (4)). Consider, for example, statement (v). From Definition 5(ii), the result $L[p(V|Y \cup U), p(Y|V \cup U)]$ is a kernel on the variable pair $(V|U)$, so that from Definition 5(i), the result $M[p(Y|V \cup U), L[p(V|Y \cup U), p(Y|V \cup U)]]$ is a kernel on the variable pair $(Y \cup V|U)$. This is the same variable pair as that for the kernel $p(Y \cup V|U)$, so that from the mentioned one-to-one correspondence, that result must be equal to this kernel. In turn, according to Equation (1), $p(Y \cup V|U)$ is equal to $p(Y|V \cup U) \vee p(V|Y \cup U)$, which is the join of the two input kernels on the left-hand side of the equation. □

As a point that is relevant to the following, let us consider this task: for any full density $p(T)$, find a (preferably small) set of kernels that are dominated by $p(T)$ and that form a *sufficient basis* for the univocal recovery of $p(T)$, using available operations on kernels. As they are all dominated by $p(T)$, the kernels in any such sufficient basis are compatible with one another. We shall see that the stated task is related to the problem of compatibility among kernels, and especially to the possible uniqueness of a consensus density for compatible kernels. The next lemma presents two exemplary answers to the task that describe sufficient bases of different forms, one of which follows a "cumulative scheme" and the other an "alternating scheme" as regards the variable pairs in the kernels.

Lemma 3. *Let $p(T)$ be a full density and $\{Y_1, \dots, Y_m\}$ be a partition of the full variable T.*

(i) *Suppose $X_1 = \varnothing$ and $X_i = Y_1 \cup \dots \cup Y_{i-1}$ for $i = 2, \dots, m$. Then the set $\{p(Y_1|X_1), \dots, p(Y_m|X_m)\}$ of kernels dominated by $p(T)$ is a sufficient basis for the recovery of $p(T)$.*

(ii) *Suppose that* $X_i = T \setminus Y_i = Y_1 \cup \cdots \cup Y_{i-1} \cup Y_{i+1} \cup \cdots \cup Y_m$ *for* $i = 1, \ldots, m$ *and that* $S(p(T)) = S(p(Y_1)) \times \cdots \times S(p(Y_m))$ *(factorability of the support for the density* $p(T)$*). Then the set* $\{p(Y_1|X_1), \ldots, p(Y_m|X_m)\}$ *of kernels dominated by* $p(T)$ *is a sufficient basis for the recovery of* $p(T)$.

Proof. (i) Consider this sequence of densities, all deducible (by projection) from the full density in question:

$$p(Y_1), \ldots, p(Y_1 \cup \cdots \cup Y_{i-1}), p(Y_1 \cup \cdots \cup Y_{i-1} \cup Y_i), \ldots, p(Y_1 \cup \cdots \cup Y_m).$$

The density $p(Y_1)$ is equal to $p(Y_1|\varnothing) = p(Y_1|X_1)$, which is available in the assumed set of kernels. Consider any $1 < i \leq m$ and suppose (as an inductive hypothesis) that the density $p(Y_1 \cup \cdots \cup Y_{i-1})$ is uniquely determined by the available kernels. Then, $p(Y_1 \cup \cdots \cup Y_{i-1} \cup Y_i)$ is also uniquely determined, since

$$p(Y_1 \cup \cdots \cup Y_{i-1} \cup Y_i) = M[p(Y_i|Y_1 \cup \cdots \cup Y_{i-1}), p(Y_1 \cup \cdots \cup Y_{i-1})]$$

by Definition 5(i) and $p(Y_i|Y_1 \cup \cdots \cup Y_{i-1}) = p(Y_i|X_i)$ is one of the available kernels. In particular, we can then conclude for $i = m$ that the full density $p(T) = p(Y_1 \cup \cdots \cup Y_m)$ is univocally recoverable (through iterated promotion) from the available kernels.

(ii) For each $i = 1, \ldots, m$, let us denote by Z_i the variable $Y_1 \cup \cdots \cup Y_i$, and then consider this sequence of saturated kernels, which are all deducible from $p(T)$ by conditioning:

$$p(Z_1|T \setminus Z_1), \ldots, p(Z_{i-1}|T \setminus Z_{i-1}), p(Z_i|T \setminus Z_i), \ldots, p(Z_m|T \setminus Z_m).$$

The first member equals $p(Y_1|T \setminus Y_1)$, which is one of the available kernels. We then consider any $1 < i \leq m$, and assume (as an inductive hypothesis) that the kernel $p(Z_{i-1}|T \setminus Z_{i-1})$ is uniquely determined by the available kernels. According to Lemma 2(v), and since $Z_i = Z_{i-1} \cup Y_i$, the following equation is true:

$$p(Z_i|T \setminus Z_i) = M[p(Z_{i-1}|T \setminus Z_{i-1}), L[p(Y_i|T \setminus Y_i), p(Z_{i-1}|T \setminus Z_{i-1})]].$$

Hence, from the inductive hypothesis concerning $p(Z_{i-1}|T \setminus Z_{i-1})$ and the fact that $p(Y_i|T \setminus Y_i)$ is one of the available kernels, we can deduce that $p(Z_i|T \setminus Z_i)$ is uniquely determined by the available kernels. In particular, we then have for $i = m$ that the full density $p(T) = p(T|\varnothing) = p(Z_m|T \setminus Z_m)$ is univocally recoverable (through an iterated lightening-and-promotion operation) from the available kernels. Note that this property can also be proved using the odds-product method that is discussed in Section 5. □

The partition hypothesized in Lemma 3 could be the subdivision $\{\{T_1\}, \ldots, \{T_n\}\}$ of the full variable $T = \{T_1, \ldots, T_n\}$ into singletons, meaning that all kernels mentioned in the lemma would be elementary kernels, that is, atoms in a lattice $\widetilde{\mathcal{P}}(T)$. Part (ii) of the lemma would then state that a full density $p(T)$ is unambiguously recoverable from the corresponding set $\{p(T_1|T \setminus \{T_1\}), \ldots, p(T_n|T \setminus \{T_n\})\}$ of elementary saturated kernels, which is a well-known result in the theory of conditionally specified probabilistic models (Besag, 1974, pp. 195–196). In addition to corollaries, Lemma 3 also admits significant generalizations. It is proven, for example, that if $\{Y_1, \ldots, Y_m\}$ is a partition of T and X_i includes (but does not necessarily equal) $Y_1 \cup \cdots \cup Y_{i-1}$ for all $i = 1, \ldots, m$, then $\{p(Y_1|X_1), \ldots, p(Y_m|X_m)\}$ is a sufficient basis for the recovery of $p(T)$ (Gelman and Speed, 1993). Furthermore, if given variables $Y_1, \ldots, Y_m$ cover the whole of T (but may overlap with one another), then the set $\{p(Y_1|T \setminus Y_1), \ldots, p(Y_m|T \setminus Y_m)\}$ of saturated kernels is a sufficient basis for the recovery of $p(T)$ (see the "if" part of Theorem 3 in Section 5). Lastly, the conditions for the recovery of a full density $p(T)$ are easily adaptable to the recovery of any kernel $p(Y|X)$. For example, if $\{Z_1, \ldots, Z_k\}$ is any partition of the variable Y, then $p(Y|X)$ is unambiguously recoverable both from the set $\{p(Z_1|X), p(Z_2|Z_1 \cup X), \ldots, p(Z_k|Z_{k-1} \cup \cdots \cup Z_1 \cup X)\}$ and from the set $\{p(Z_1|(Y \setminus Z_1) \cup X), \ldots, p(Z_k|(Y \setminus Z_k) \cup X)\}$ of kernels dominated by it in a lattice.

3. Compatibility of probability kernels and sure compatibility of variable pairs

The algebraic frame outlined in the preceding section allows us to assign a suitable place to the main concept of our study.

Definition 6. Given probability kernels $p(Y_1|X_1), \ldots, p(Y_m|X_m)$ on variable pairs within a full variable T are *compatible* with one another if there exists a full density $p(T)$ that dominates each of them. Such $p(T)$ is said to be a *consensus density* for the given kernels.

In other words, compatibility within a set of kernels means that there exists a lattice of kernels that includes that set. This concept is a topic in the literature concerning conditionally specified probabilistic models (Arnold, Castillo and Sarabia, 1999, p. 5; Kaiser, 2002, p. 1213). Terms such as "candidate, putative, proposed conditionals" are used in relation to kernels whose mutual compatibility is under judgement (Arnold, Castillo and Sarabia, 2004, pp. 147, 157).

As a first comment on the above definition, we remark that there are alternative ways of expressing the compatibility relation. In particular, we can refer to the join $(Z|W) = (Y_1|X_1) \vee \cdots \vee (Y_m|X_m)$ of the variable pairs in the candidate kernels, and state that these are compatible if there exists a kernel $p(Z|W)$ that dominates each of them. As a second comment, we note that a uniqueness problem and a construction problem are naturally associated with the compatibility problem (concerning existence). That is, if there are reasons for stating that given kernels are compatible, then one may ask whether there is only one consensus density for them, or a (possibly infinite) number of such densities, and look for practical procedures for discovering or constructing such a density. The concept of a "sufficient basis", which was discussed at the end of the preceding section, is related to these problems. As a third comment, we remark that compatibility is not a transitive relation in general. For example, if X and Y are disjoint variables, then any density $p(X)$ is compatible both with any density $p(Y)$ and with any kernel $p(Y|X)$, although these two may fail to be compatible unless $p(Y) = J[M[p(Y|X), p(X)], X]$ (see Lemma 2(i)). This lack of transitivity implies that compatibility within a set of three or more kernels cannot be approved solely on the basis of pairwise compatibility, and is a sign of the difficulty of the problem. Indeed, the compatibility problem may become quite tricky, as revealed by paradoxical situations noted in the literature. It is surprising, for example, that if $p(Y|X)$ and $p(X|Y)$ are deduced (by conditioning) from a certain density $p(X \cup Y)$, a collection $\mathcal{Q}$ of kernels may nevertheless exist such that compatibility is true within $\mathcal{Q} \cup \{p(Y|X), p(X|Y)\}$ but false within $\mathcal{Q} \cup \{p(X \cup Y)\}$, even though $\{p(Y|X), p(X|Y)\}$ and $p(X \cup Y)$ are substantially equivalent, since the latter (under suitable conditions) may be recovered from the former through lightening-and-promotion, as shown by Lemma 2(v) (Kuo and Wang, 2011, pp. 2460–2461).

The next definition focuses on variable pairs as the set-theoretic carriers of probability kernels.

Definition 7. A set $\{(Y_1|X_1), \dots, (Y_m|X_m)\}$ of variable pairs has *sure compatibility* if every set

$$\{p(Y_1|X_1), \dots, p(Y_m|X_m)\}$$

of kernels definable on those pairs satisfies the compatibility condition.

Note that given any set $\{(Y_1|X_1), \dots, (Y_m|X_m)\}$ of variable pairs in a full variable T, there is certainly *some* set $\{p(Y_1|X_1), \dots, p(Y_m|X_m)\}$ of kernels on those pairs that are mutually compatible: we can simply consider any density $p(T)$ on the full variable and then derive the kernels from it by suitable projections and/or conditionings. However, Definition 7 demands much more than this: it demands that *every* possible set $\{p(Y_1|X_1), \dots, p(Y_m|X_m)\}$ of kernels on the given variable pairs satisfies compatibility, meaning that the root of compatibility is to be found not in the numerical characteristics of the particular kernels considered but in the set-theoretic characteristics of the variable pairs themselves. For example, if X and Y are disjoint variables, then the set $\{(X|\varnothing), (Y|\varnothing)\}$ has sure compatibility, since for any densities $p(X) = p(X|\varnothing)$ and $q(Y) = q(Y|\varnothing)$ we can consider (for example) the product density $r(X \cup Y) = p(X) \cdot q(Y)$, which dominates (by projection) both $p(X|\varnothing)$ and $q(Y|\varnothing)$, so that these two are compatible with each other. Conversely, if X and Y are not disjoint, then any given densities $p(X) = p(X|\varnothing)$ and $q(Y) = q(Y|\varnothing)$ are compatible only if $p(X \cap Y) = q(X \cap Y)$ (with $p(X \cap Y) = J[p(X), X \setminus Y]$ and $q(X \cap Y) = J[q(Y), Y \setminus X]$), so that the set $\{(X|\varnothing), (Y|\varnothing)\}$ has not sure compatibility.

The next lemma highlights two further counterexamples to sure compatibility.

Lemma 4. (i) *For all variable pairs* $(V|U)$ *and* $(Y|X)$ *such that* $V \cap Y \neq \varnothing$, *the set* $\{(V|U), (Y|X)\}$ *has not sure compatibility.*

(ii) *Any circular set* $\{(Z_m|Z_{m-1}), (Z_{m-1}|Z_{m-2}), \dots, (Z_2|Z_1), (Z_1|Z_m)\}$ *of non-null variable pairs has not sure compatibility.*

Proof. (i) Since $V \cap Y$ is a non-empty variable, there are densities $p(V \cap Y)$ and $q(V \cap Y)$ that are different from each other. Choose any densities $p(V)$ and $q(Y)$ such that $p(V \cap Y) \preceq p(V)$ and $q(V \cap Y) \preceq q(Y)$, and then construct the kernels $p(V|U)$ and $q(Y|X)$ by setting $p(V|u) = p(V)$ for all $u \in U^\circ$ and $q(Y|x) = q(Y)$ for all $x \in X^\circ$. We claim that these kernels are not compatible with each other (which then implies that the set $\{(V|U), (Y|X)\}$ has not sure compatibility). Indeed, suppose (ad absurdum) that they are compatible, so that there is a full density $r(T)$ such that $p(V|U) \preceq r(T)$ and $q(Y|X) \preceq r(T)$. Then, considering the way in which $p(V|U)$ and $q(Y|X)$ are constructed, we should also have $p(V) \preceq r(T)$ and $q(Y) \preceq r(T)$ by Lemma 1, and then $p(V \cap Y) \preceq r(T)$ and $q(V \cap Y) \preceq r(T)$ by transitivity. However, this is impossible, because $p(V \cap Y)$ and $q(V \cap Y)$ act on the same variable and are different.

(ii) Given a circular set $\{(Z_m|Z_{m-1}), \dots, (Z_2|Z_1), (Z_1|Z_m)\}$ of variable pairs, let us construct a corresponding set $\{p_{m-1}(Z_m|Z_{m-1}), \dots, p_1(Z_2|Z_1), p_m(Z_1|Z_m)\}$ of kernels with these characteristics:

for each $i = 1, \dots, m-1$, the kernel $p_i(Z_{i+1}|Z_i)$ is of deterministic type,
which means that there is a function f_i from Z_i° to Z_{i+1}° such that
$p_i(v|u) = 1$ or $= 0$, depending on whether v equals or does not equal $f_i(u)$,
for all $u \in Z_i^\circ$ and $v \in Z_{i+1}^\circ$;

the kernel $p_m(Z_1|Z_m)$ is such that
there are $t \in Z_1^\circ$ and $w \neq w' \in Z_m^\circ$ with $p_m(t|w) > 0$ and $p_m(t|w') > 0$.

Our claim is that such kernels are not mutually compatible (which then implies that the given circular set of variable pairs has not sure compatibility). Suppose the contrary, that is, the existence of a consensus density $r(T)$, so that

$$r(Z_{i+1}|Z_i) = p_i(Z_{i+1}|Z_i) \text{ for all } i = 1, \dots, m-1 \quad \text{ and } \quad r(Z_1|Z_m) = p_m(Z_1|Z_m).$$

It can be seen that, due to the deterministic character of the kernels $r(Z_m|Z_{m-1}), \dots, r(Z_2|Z_1)$, under the density $r(T)$ for each point $u \in Z_1^\circ$ there must be a single point $g(u) \in Z_m^\circ$ with positive conditional probability given the hypothesis $Z_1 = u$. In other words, the kernel $r(Z_m|Z_1)$ that is deducible from $r(T)$ must itself be deterministic. In regard to the above-mentioned points $t \in Z_1^\circ$ and $w \neq w' \in Z_m^\circ$, this implies, in particular, that we cannot have both $r(t, w) > 0$ and $r(t, w') > 0$, whereas from the construction of $p_m(Z_1|Z_m)$ and the equality $r(Z_1|Z_m) = p_m(Z_1|Z_m)$ we should have both $r(t, w) > 0$ and $r(t, w') > 0$, which is a contradiction. Therefore, the constructed kernels cannot have a consensus density; that is, they are not mutually compatible. □

In the next paragraph, we will present a characterization of the concept of "sure compatibility". In expressing the characterization, use will be made of a simple binary relation between variable pairs that is different from the dominance relation specified in Definition 1.

Definition 8. A variable pair $(V|U)$ is *incident* on a variable pair $(Y|X)$ (notation $(V|U) \to (Y|X)$) if $V \cap X \neq \varnothing$.

This relation is areflexive, simply because in any variable pair the two variables are assumed to be disjoint. Besides, it is free as regards other possible formal properties of binary relations, such as symmetry, asymmetry, transitivity, acyclicity, and so on. For example, if X and Y are disjoint non-empty variables, then the set $\{(Y|X), (X|Y)\}$ of two symmetric variable pairs forms a cycle of length two according to incidence. If $T = \{T_1, T_2, T_3, T_4, T_5\}$, then the pairs $\{(T_1, T_3|T_4, T_5), (T_4|T_2), (T_2, T_3|T_1)\}$ form a cycle of length three, whereas within the set $\{(T_5|T_3, T_4), (T_4|T_1, T_2), (T_3|T_1), (T_1|\varnothing)\}$ the incidence relation is acyclic.

The following is a salient result of our discussion of compatibility in this article.

Theorem 1. *A set* $\{(Y_1|X_1), \dots, (Y_m|X_m)\}$ *of variable pairs in a full variable* T *has sure compatibility of kernels if and only if* (i) *the conditioned variables* $Y_1, \dots, Y_m$ *in the pairs are disjoint from one another and* (ii) *the incidence* $\to$ *between the pairs is an acyclic relation.*

Proof. "Only if" part. The necessity of condition (i) follows directly from Lemma 4(i), since if a set of variable pairs has sure compatibility, then each of its subsets must also have this property. To prove the necessity of condition (ii), let us first suppose that the given set of variable pairs forms a →-cycle, specifically

$$(Y_m|X_m) \to (Y_1|X_1) \to (Y_2|X_2) \to \cdots \to (Y_{m-2}|X_{m-2}) \to (Y_{m-1}|X_{m-1}) \to (Y_m|X_m),$$

which means that the following variables are all non-empty

$$Z_1 = X_1 \cap Y_m,\ \ Z_2 = X_2 \cap Y_1, \dots, Z_{m-1} = X_{m-1} \cap Y_{m-2},\ \ Z_m = X_m \cap Y_{m-1}.$$

Thus, $\{(Z_m|Z_{m-1}), \dots, (Z_2|Z_1), (Z_1|Z_m)\}$ is a circular set of non-null variable pairs and Lemma 4(ii) ensures the existence of a set of kernels $\{p_{m-1}(Z_m|Z_{m-1}), \dots, p_1(Z_2|Z_1),\ p_m(Z_1|Z_m)\}$ that are not mutually compatible. For each $i = 1, \dots, m$ (with $i-1 = m$ for $i = 1$, and $i+1 = 1$ for $i = m$), we can expand the kernel

$$p_i(Z_{i+1}|Z_i) = p_i(X_{i+1} \cap Y_i|X_i \cap Y_{i-1})$$

into a kernel

$$p_i(Y_i|X_i) = p_i(Z_{i+1} \cup (Y_i \setminus X_{i+1})|Z_i \cup (X_i \setminus Y_{i-1}))$$

by first constructing

$$p_i(Z_{i+1} \cup (Y_i \setminus X_{i+1})|Z_i) = M[p_i(Y_i \setminus X_{i+1}|Z_{i+1} \cup Z_i),\ p_i(Z_{i+1}|Z_i)]$$

where $p_i(Y_i \setminus X_{i+1}|Z_{i+1} \cup Z_i)$ is an arbitrarily chosen kernel, and then setting

$$p_i(Z_{i+1} \cup (Y_i \setminus X_{i+1})|Z_i, u) = p_i(Z_{i+1} \cup (Y_i \setminus X_{i+1})|Z_i) \text{ for every } u \in (X_i \setminus Y_{i-1})^\circ.$$

It can be seen that if we have a full density $p(T)$ such that $p_i(Y_i|X_i) \preceq p(T)$ for all $i = 1, \dots, m$, then, from Lemmas 1 and 2(iii), we also have $p_i(Z_{i+1}|Z_i) \preceq p(T)$ for all $i = 1, \dots, m$, which contradicts the assumption that the kernels $\{p_i(Z_{i+1}|Z_i)\}$ are not compatible. Hence, the kernels $\{p_i(Y_i|X_i)\}$ constructed in this way are not compatible, which proves that the set of pairs $\{(Y_i|X_i)\}$ has not sure compatibility. Lastly, if the set of pairs $\{(Y_1|X_1), \dots, (Y_m|X_m)\}$ were not a →-cycle but included a subset that formed a →-cycle, then the argument developed above could be applied to that subset, thus showing that not only the subset but also the entire set including it has not sure compatibility.

"If" part. Suppose that $\{(Y_1|X_1), \dots, (Y_m|X_m)\}$ is a set of variable pairs that complies with conditions (i) and (ii) in the theorem. Property (ii) implies that there is a permutation $((Y_{s(1)}|X_{s(1)}), \dots, (Y_{s(m)}|X_{s(m)}))$ of the given set such that not$((Y_{s(i)}|X_{s(i)}) \to (Y_{s(j)}|X_{s(j)}))$ for all $1 \le j < i \le m$. Together with property (i), this implies

$$Y_{s(i)} \cap (Y_{s(i-1)} \cup X_{s(i-1)} \cup \cdots \cup Y_{s(1)} \cup X_{s(1)}) = \varnothing \quad \text{for all } i = 2, \dots, m. \tag{7}$$

The proof of sure compatibility is by induction on the number $m \geq 2$ of variable pairs.

First step: For $m = 2$, let any two variable pairs $(Y_{s(1)}|X_{s(1)}) = (Y|X)$ and $(Y_{s(2)}|X_{s(2)}) = (V|U)$ be given such that $V \cap (Y \cup X) = \varnothing$, that is $FG = \varnothing$ in the terms of Figure 1, so that $(Y|X) \vee (V|U) = (ABCDE|H)$ according to Equation (1) (here and in the rest of this proof the symbol $\cup$ is omitted for simplicity, so that FG and $ABCDE$ stand for $F \cup G$ and $A \cup B \cup C \cup D \cup E$, respectively). Let $p(Y|X) = p(AE|CH)$ and $p(V|U) = p(B|DEH)$ be arbitrary kernels on the variable pairs. First, we extend $p(B|DEH)$ into $p(B|ACDEH)$ by setting

$$p(B|a, c, DEH) = p(B|DEH), \text{ for all } (a, c) \in A^\circ \times C^\circ. \tag{8}$$

Then by multiple promotion we can construct this kernel

$$p(ABCDE|H) = p(B|ACDEH) \ M \ p(D|ACEH) \ M \ p(AE|CH) \ M \ p(C|H)$$

where $p(D|ACEH)$ and $p(C|H)$ are freely chosen kernels (this writing takes account of the associativity of the promotion operation, as noted in Equation (6)). The kernel $p(ABCDE|H)$ thus constructed dominates $p(AE|CH)$ due to parts (ii) and (iii) of Lemma 2, and dominates $p(B|ACDEH)$ due to part (ii) of that lemma. Hence, it also dominates $p(B|DEH)$ on account of Equation (8) and of Lemma 1. The kernels $p(Y|X) = p(AE|CH)$ and $p(V|U) = p(B|DEH)$ (which are arbitrary) are therefore compatible with each other, as there is a kernel $p(ABCDE|H)$ on the join variable pair $(Y|X) \vee (V|U) = (ABCDE|H)$ that dominates both of them. Note that $p(ABCDE|H)$ does not involve any variables besides those involved in $p(AE|CH)$ or in $p(B|DEH)$.

Inductive step: Let us now consider any list

$$(p_{s(1)}(Y_{s(1)}|X_{s(1)}), \ldots, p_{s(m-1)}(Y_{s(m-1)}|X_{s(m-1)}), p_{s(m)}(Y_{s(m)}|X_{s(m)}))$$

of $m > 2$ kernels whose variable pairs comply with condition (7), and suppose (as an inductive hypothesis) that the first $m - 1$ members in the list are compatible with one another, so that there is a kernel $p(Z|W)$ that dominates all of them. Based on the remark that concludes the preceding step in the current proof, we may presume that $Z \cup W \subseteq Y_{s(m-1)} \cup X_{s(m-1)} \cup \cdots \cup Y_{s(1)} \cup X_{s(1)}$, so that $Y_{s(m)} \cap (Z \cup W) = \varnothing$ due to hypothesis (7). Thus, the conditions are satisfied that make it possible to apply the argument in the preceding step to the kernels $p(Z|W)$ and $p_{s(m)}(Y_{s(m)}|X_{s(m)})$. This argument ensures the existence of a kernel that dominates both $p(Z|W)$ (and hence

$$p_{s(1)}(Y_{s(1)}|X_{s(1)}), \ldots, p_{s(m-1)}(Y_{s(m-1)}|X_{s(m-1)})$$

by transitivity) and $p_{s(m)}(Y_{s(m)}|X_{s(m)})$. The m kernels are therefore all compatible with one another. As these are arbitrary, this proves that the given set of variable pairs has sure compatibility. □

The theorem thus proved characterizes the sure compatibility of a set of variable pairs by referring only to the set-theoretic properties of those pairs, rather than to the numerical properties of the probability kernels definable on them. The "if" part of the theorem ensures that if a given set of pairs satisfies the set-theoretic conditions (i) and (ii), then no further test is required to accept the compatibility hypothesis of any set of kernels acting on those pairs. The "only if" part signifies, complementarily, that if conditions (i) and (ii) are not both satisfied, then the acceptance (or refusal) of the compatibility hypothesis requires further tests of the numerical characteristics of the kernels under judgement. Let us consider, for example, the two schemes presented in Lemma 3. We can see that the cumulative scheme (e.g., $\{(Y_1|\varnothing), (Y_2|Y_1), (Y_3|Y_1 \cup Y_2)\}$ for $m = 3$) complies with conditions (i) and (ii), so that arbitrary kernels defined on the variable pairs in the scheme are mutually compatible. From Lemma 3(i), there is a single consensus density for the given kernels, which can be constructed from these by multiple promotion. On the contrary, the alternating scheme (e.g., $\{(Y_1|Y_2 \cup Y_3), (Y_2|Y_1 \cup Y_3), (Y_3|Y_1 \cup Y_2)\}$ for $m = 3$) violates condition (ii) (indeed, the incidence relation within that scheme forms a complete directed graph, thus containing cycles) so that compatibility is not generally ensured for kernels definable on the variable pairs in the scheme. From Lemma 3(ii), if kernels defined on the variable pairs in an alternating scheme admit a consensus density whose support is factorable, then this density is unique and can be constructed from the given kernels through lightening-and-promotion operations.

4. Compatibility beyond structural assurance: the two kernels case

In light of the preceding discussion, we can expect that the situations explored in research on the compatibility of distributions are those in which the two conditions in Theorem 1 are not both satisfied, so that compatibility is not structurally guaranteed. The simplest of these situations involves a set $\{p(Y|X), q(X|Y)\}$ of two kernels on *symmetric variable pairs*. The pairs $(Y|X)$ and $(X|Y)$ form a cycle (of length two) according to the incidence relation (Definition 8), and thus they falsify condition (ii) of Theorem 1. In the first half of this section, we review some results from the literature that characterize compatibility within a pair of kernels on symmetric variable pairs. We review them from the standpoint defined in the preceding section and, for simplicity, limit ourselves to results applicable to kernels $p(Y|X)$ and $q(X|Y)$ that satisfy the following *positivity condition* (see footnote 3):

$$p(y|x) > 0 \text{ and } q(x|y) > 0 \text{ for all } (x, y) \in X^\circ \times Y^\circ. \tag{9}$$

In the literature, however, there are also generalizations of these results that apply to kernels satisfying the following, less restrictive condition:

$$p(y|x) > 0 \text{ if and only if } q(x|y) > 0 \text{ for all } (x, y) \in X^\circ \times Y^\circ. \tag{10}$$

Note that this condition is necessary for compatibility, because if kernels $p(Y|X)$ and $q(X|Y)$ are both dominated by a density $r(X \cup Y)$ such that $S(r(X)) = X^\circ$ and $S(r(Y)) = Y^\circ$, then, according to Definition 3(ii), $p(y|x) = r(x, y)/r(x)$ and $q(x|y) = r(x, y)/r(y)$ for all $(x, y) \in X^\circ \times Y^\circ$, so that $p(y|x)$ and $q(x|y)$ are positive precisely when $r(x, y)$ is positive. In the second half of the current section, we will present a result of our own analysis, which characterizes compatibility between kernels on arbitrary variable pairs $(Y|X)$ and $(V|U)$, thus going beyond the special case of symmetric variable pairs.

One characterization is expressed by the following statement (Arnold and Press, 1989).

Proposition 2. *Two kernels $p(Y|X)$ and $q(X|Y)$ on symmetric variable pairs are compatible if and only if there are densities $p(X)$ and $q(Y)$ such that $p(y|x) \cdot p(x) = q(x|y) \cdot q(y)$ for all $(x, y) \in X^\circ \times Y^\circ$.*

In the notation used in Definition 5(i), the equation in this proposition can be rewritten as

$$M[p(Y|X), p(X)] = M[q(X|Y), q(Y)]. \tag{11}$$

The truth of the proposition is clarified by noting that the existence of densities $p(X)$ and $q(Y)$ that satisfy equation (11) is tantamount to the existence of a common upper bound $p(X \cup Y) = q(X \cup Y)$ for $p(Y|X)$ and $q(X|Y)$ within a lattice of kernels, which is precisely the meaning of compatibility between the given kernels. The stated characterization is of *existential* type, as it links the compatibility between $p(Y|X)$ and $q(X|Y)$ to the existence of a solution to Equation (11) in the unknowns $p(X)$ and $q(Y)$.

A second characterization involves two functions that are deducible from the kernels under consideration. Specifically, once a reference point $(x', y') \in X^\circ \times Y^\circ$ has been arbitrarily chosen, two functions $f(X, Y)$ and $g(X, Y)$ can be separately derived from the kernels $p(Y|X)$ and $q(X|Y)$ by setting, for each $(x, y) \in X^\circ \times Y^\circ$,

$$f(x, y) = \frac{p(y|x) \cdot p(y'|x')}{p(y'|x) \cdot p(y|x')}, \quad g(x, y) = \frac{q(x|y) \cdot q(x'|y')}{q(x'|y) \cdot q(x|y')}.$$

In other words, $f(X, Y)$ is obtained as an odd-ratio function based on $p(Y|X)$, and similarly $g(X, Y)$ from $q(X|Y)$ (all ratios exist as real numbers, under the positivity condition (9)). In these terms, the following relationship is true (Arnold and Press, 1989, p. 52; Chen, 2010, p. 672).

Proposition 3. *Kernels $p(Y|X)$ and $q(X|Y)$ are compatible if and only if f(X,Y)=g(X,Y).*

The truth of the "only if" part is easily seen: if $p(Y|X)$ and $q(X|Y)$ are dominated by the same density $r(X \cup Y)$, then $p(y|x) = r(x, y)/r(x)$ and $q(x|y) = r(x, y)/r(y)$, so that $f(x, y) = r(x, y) \cdot r(x', y')/r(x', y) \cdot r(x, y') = g(x, y)$ for all $(x, y) \in X^\circ \times Y^\circ$. The given characterization is of *deductive* type: it expresses compatibility in terms of the equality between two functions $f(X, Y)$ and $g(X, Y)$ that are deducible from $p(Y|X)$ and $q(X|Y)$ in the described way.

A third characterization involves another function that is deducible from the kernels in question. Specifically, based on $p(Y|X)$ and $q(X|Y)$, a ratio function $h(X, Y)$ can be constructed by setting, for all $(x, y) \in X^\circ \times Y^\circ$,

$$h(x, y) = \frac{p(y|x)}{q(x|y)}$$

which again is a legitimate computation under the positivity condition. The following statement holds true (Arnold and Press, 1989, pp. 152–153; Tian, Tan, Ng and Tang, 2009, p. 119):

Proposition 4. *Kernels $p(Y|X)$ and $q(X|Y)$ are compatible if and only if there are functions $a(X)$ and $b(Y)$ such that $h(x, y) = a(x) \cdot b(y)$ for all $(x, y) \in X^\circ \times Y^\circ$.*

As above, the "only if" part is easily proved: if $p(Y|X)$ and $q(X|Y)$ are dominated by the same density $r(X \cup Y)$, then $h(x, y) = p(y|x)/q(x|y) = (r(x, y)/r(x))/(r(x, y)/r(y)) = (1/r(x)) \cdot r(y)$ for all $(x, y) \in X^\circ \times Y^\circ$, so that by setting $a(X) = 1/r(X)$ and $b(Y) = r(Y)$ a factorization of $h(X, Y)$ in the asserted form is obtained.

In particular, if sets X° and Y° have finite cardinality, then the function $h(X, Y)$ can be represented as a matrix of $|X^\circ|$ rows and $|Y^\circ|$ columns. The equation $h(X, Y) = a(X) \cdot b(Y)$ would then mean that this matrix is expressible as the product of a column vector $a(X)$ by a row vector $b(Y)$, implying that all rows in the matrix are proportional to one another (and similarly for the columns). As a consequence, when referring to variables with finite sets of possible values, the above connection can be reformulated as follows (Arnold, Castillo and Sarabia, 2004, p. 137; Kuo, Song and Jiang, 2017, pp. 117–118).

Proposition 5. *Kernels $p(Y|X)$ and $q(X|Y)$ are compatible if and only if the matrix $h(X, Y)$ has rank* 1.

Note that, although the general characterization in Proposition 4 is of the existential type (it demands the existence of functions $a(X)$ and $b(Y)$ satisfying a definite equation), the specific version in Proposition 5 is of the deductive type, as it concerns a possible property (unit rank) of the matrix $h(X, Y)$ that is deducible from the given kernels. We also remark that, besides this elementary result, there are other ways in which linear algebra and associated geometrical arguments have proved of use in discussing compatibility of distributions. For example, Arnold, Castillo, and Sarabia (2002, pp. 235–239) on considering any pair of kernels $\{p(Y|X), q(X|Y)\}$ that comply with (10) but may violate (9), show how the search for a consensus distribution $r(X \cup Y)$ can be formalized as the task of solving a definite system of linear equations and inequalities. Thus, the set of possible solutions is tantamount to a convex subset of a geometric space and may be explored using methods of linear programming.

The characterizations reviewed so far are limited in scope, as they apply only to any pair of kernels $\{p(Y|X), q(X|Y)\}$ on symmetric variable pairs. With the next theorem, we contribute to the topic of compatibility by presenting a characterization that is applicable to any pair of kernels $\{p(V|U), q(X|Y)\}$ that are free of constraints on the variable pairs. In stating and proving this theorem, use will be made of the set-theoretic labeling represented in Figure 1 and the basic operations on kernels defined in Section 2. As in the proof of Theorem 1, the symbol $\cup$ will be omitted for brevity when specifying composite variables (e.g., BFG stands for $B \cup F \cup G$).

Theorem 2. *Any two kernels $p(V|U) = p(BFG|DEH)$ and $q(Y|X) = q(AEF|CGH)$ on variable pairs in a full variable T are compatible with each other if and only if there are kernels $p(DE|H)$ and $q(CG|H)$ that satisfy the equation*

$$J[M[p(BFG|DEH), p(DE|H)], BD] = J[M[q(AEF|CGH), q(CG|H)], AC], \tag{12}$$

where J and M are the projection and promotion operations.

Proof. Only if" part. If the given kernels are compatible, then there is a consensus full density $r(T)$ for them, such that $p(BFG|DEH) = r(BFG|DEH)$ and $q(AEF|CGH) = r(AEF|CGH)$. By setting $p(DE|H) = r(DE|H)$ and $q(CG|H) = r(CG|H)$, we find that both sides of Equation (12) specify the kernel $r(EFG|H)$, so that the equation is satisfied.

"If" part. Let $p(V|U) = p(BFG|DEH)$ and $q(Y|X) = q(AEF|CGH)$ be arbitrary kernels on the indicated variable pairs, and suppose that there are kernels $p(DE|H)$ and $q(CG|H)$ that when combined with them satisfy Equation (12). By promotion, we can first construct the kernels

$$p(BDEFG|H) = M[p(BFG|DEH), p(DE|H)], \tag{13}$$

$$q(ACEFG|H) = M[q(AFE|CGH), q(CG|H)], \tag{14}$$

from which we may derive the following further kernels by conditioning

$$p(BD|EFGH) = C[p(BDEFG|H), EFG], \tag{15}$$

$$q(AC|EFGH) = C[q(ACEFG|H), EFG]. \tag{16}$$

Hypothesis (12) means that the projection (relative to BD) of the kernel defined in (13) equals the projection (relative to AC) of the kernel defined in (14), so that the same symbol $r(EFG|H)$ may be used for both projections:

$$r(EFG|H) = J[p(BDEFG|H), BD] = J[q(ACEFG|H), AC]. \tag{17}$$

Furthermore, the kernels defined in (15) and (16) act on variable pairs that have the same conditioning variable $EFGH$ and disjoint conditioned variables BD and AC. Thus, by Theorem 1 the pair of such variable pairs has sure compatibility, implying that there exists some kernel $r(ABCD|EFGH)$ that dominates both kernels. More precisely

$$p(BD|EFGH) = J[r(ABCD|EFGH), AC], \tag{18}$$

$$q(AC|EFGH) = J[r(ABCD|EFGH), BD]. \tag{19}$$

Lastly, the kernels $r(ABCD|EFGH)$ and $r(EFG|H)$ are suitable for promotion, thus producing the result

$$r(ABCDEFG|H) = M[r(ABCD|EFGH), r(EFG|H)]. \tag{20}$$

We now prove that this kernel (which acts on the join variable pair $(ABCDEFG|H) = (V|U) \vee (Y|X)$) dominates both $p(V|U)$ and $q(Y|X)$, so that these are compatible. Indeed:

$$\begin{aligned}
&C[J[r(ABCDEFG|H), AC], DE] = \text{ by (20)}\\
&C[J[M[r(ABCD|EFGH), r(EFG|H)], AC], DE] = \text{ by Lemma 2(iv)}\\
&C[M[J[r(ABCD|EFGH), AC], r(EFG|H)], DE] = \text{ by (18)}\\
&C[M[p(BD|EFGH), r(EFG|H)], DE] = \text{ by (15) and (17)}\\
&C[M[C[p(BDEFG|H), EFG], J[p(BDEFG|H), BD]], DE] = \text{ by Lemma 2(vi)}\\
&C[p(BDEFG|H), DE] = \text{ by Definition 3(ii)}\\
&p(BFG|DEH) = p(V|U),
\end{aligned}$$

so that $r(ABCDEFG|H)$ dominates $p(V|U)$. The dominance of $r(ABCDEFG|H)$ over $q(Y|X)$ is proved by a similar argument. □

The characterization in Theorem 2 is of the existential type, as it demands the existence (and the discovery in actual applications) of kernels $p(DE|H)$ and $q(CG|H)$ such that when combined with the given kernels $p(V|U)$ and $q(Y|X)$ through promotion-and-projection, Equation (12) is verified. From Theorem 2, several corollaries may be deduced by setting constraints on the kernels $p(V|U)$ and $q(Y|X)$

under consideration, more precisely by assuming that certain parts of the variables they involve are empty. For example, if $ABCDFH = \varnothing$, then Equation (12) becomes

$$J[M[p(G|E),\, p(E|\varnothing)],\, \varnothing] = J[M[q(E|G),\, q(G|\varnothing)],\, \varnothing],$$

that is, due to (3),

$$M[p(G|E),\, p(E)] = M[q(E|G),\, q(G)],$$

which is a rewriting of (11). Hence, the characterization in Proposition 2 amounts to a special case of the characterization in Theorem 2. As another example, if $CDEG = \varnothing$, then Equation (12) becomes

$$J[M[p(BF|H),\, p(\varnothing|H)],\, B] = J[M[q(AF|H),\, q(\varnothing|H)],\, A],$$

that is, due to (5),

$$J[p(BF|H),\, B] = J[q(AF|H),\, A].$$

This formally corroborates the following intuitive principle: any two kernels with the same conditioning variable and partially overlapping conditioned variables are compatible if and only if their projections on the intersection of the conditioned variables are equal.

5. Compatibility beyond structural assurance: the multiple kernels case

A natural generalization of the case discussed in the first half of the preceding section (that is, a pair of kernels $\{p(Y|X),\, q(X|Y)\}$ on symmetric variable pairs) is given by any set $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ of *saturated* kernels whose conditioned variables are *exhaustive* of the full variable T. A notable example of this is the alternating scheme represented in Lemma 3(ii), in which the conditioned variables $Y_1, \ldots, Y_m$ more precisely form a partition of T. In this section, we then refer to any set of kernels $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ whose variable pairs comply with these conditions:

$$X_i = T \setminus Y_i \text{ for each } i = 1, \ldots, m \text{ (the kernels are saturated);} \tag{21}$$

$$Y_1 \cup \cdots \cup Y_m = T \text{ (the conditioned variables are exhaustive).} \tag{22}$$

Note that within a set of variable pairs with these properties, the incidence relation → in Definition 8 can give rise to cycles (for example, within an alternating scheme, it determines a complete directed graph that obviously has cycles), so that based on Theorem 1, such a set of variable pairs could fail to have sure compatibility. In that situation, a decision concerning the compatibility between given kernels should then be taken by examining the numerical properties of the kernels themselves, as families of density functions. Here, we review an exemplary decision criterion limited to densities on finite domains, a criterion that has been variously studied in the literature.

Let $T = \{T_1, \ldots, T_n\}$ be a full variable whose elements are variables with *finite* sets of possible values. For any point t in the space T° and any sub-variable U of T, let t_U denote the *projection* of the point t

on the space U°. In formal terms:

$$\text{for any } t = (t_1, \ldots, t_n) \in T_1^\circ \times \cdots \times T_n^\circ = T^\circ \text{ and any } U = \{T_{g(1)}, \ldots, T_{g(k)}\} \subseteq T$$
$$t_U \text{ stands for } (t_{g(1)}, \ldots, t_{g(k)}).$$

For example, if $T = \{T_1, T_2, T_3, T_4, T_5\}$, $U = \{T_2, T_3, T_5\}$, and $t = (2, 3, 1, 3, 2)$, then $t_U = (3, 1, 2)$. Using this notation and referring to any set $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ of saturated kernels, we first remark that if these kernels are compatible, then the following condition must be true:

$$\text{for all } t \in T^\circ \text{ and all } 1 \leq i, j \leq m \qquad (23)$$
$$p_i(t_{Y_i}|t_{X_i}) > 0 \text{ if and only if } p_j(t_{Y_j}|t_{X_j}) > 0.$$

This condition generalizes requirement (10), and its necessity for compatibility can be proved by an argument similar to that used for that requirement. Our discussion in this section, however, is focused on sets of kernels that satisfy the *positivity condition*

$$p_i(y_i|x_i) > 0 \quad \text{for all } x_i \in X_i^\circ,\ y_i \in Y_i^\circ, \text{ and } i = 1, \ldots, m, \qquad (24)$$

which is stronger than (23) and in turn generalizes (9). After presenting the main result, in the last paragraph we will comment on the complications that may arise when the kernels comply with (23) but not with (24), that is, when there are "structural zeros" in the kernels under consideration.

In the assumed conditions, for each $i = 1, \ldots, m$ an *adjacency* relation E_i within the space T° can be determined by setting

$$E_i = \{(s, i, t) : s, t \in T^\circ, s_{X_i} = t_{X_i}\}. \qquad (25)$$

In other words, any two points $s = (s_1, \ldots, s_n)$ and $t = (t_1, \ldots, t_n)$ in the space T° are adjacent according to E_i if they coincide in all the coordinates for the elementary variables in X_i (and thus may only differ in some of the coordinates for the elementary variables in $Y_i = T \setminus X_i$). For example, suppose $T = \{T_1, \ldots, T_5\}$, $T^\circ = \{1, 2, 3\}^5$, and $X_i = \{T_2, T_3\}$, and consider the points $s = (1, 3, 1, 2, 2)$, $t = (2, 3, 1, 3, 2)$, and $u = (1, 2, 1, 2, 2)$. Then $(s, i, t) \in E_i$ because $s_{X_i} = (3, 1) = t_{X_i}$, whereas $(s, i, u) \notin E_i$ because $s_{X_i} = (3, 1) \neq (2, 1) = u_{X_i}$. Overall, we can then consider a relational structure

$$(T^\circ, E) = (T^\circ, E_1 \cup \cdots \cup E_m)$$

which formally amounts to a graph with the space T° as the set of points and the relation $E = E_1 \cup \cdots \cup E_m$ as the set of lines.

We note the following properties of this graphical structure. First, the lines in the graph are *labeled* and *directed*. Indeed, each line (s, i, t) has the label i, which indicates the adjacency E_i to which the line belongs, and it is counted as distinct from the inverse line (t, i, s), which also belongs to E_i. Secondly, each adjacency E_i has the formal properties of an *equivalence* (reflexivity, symmetry, and transitivity). However, the pooled adjacency $E = E_1 \cup \cdots \cup E_m$ may fail to be transitive because for any points s, t, and u, the existence of some E_i and E_j such that $(s, i, t) \in E_i$ and $(t, j, u) \in E_j$ does not ensure the existence of some E_h such that $(s, h, u) \in E_h$. Thirdly, any two points in T° may have *multiple* *adjacency*. For example, referring to the case mentioned in the preceding paragraph with $s = (1, 3, 1, 2, 2)$ and $t = (2, 3, 1, 3, 2)$, and assuming $X_i = \{T_2, T_3\}$ and $X_j = \{T_3, T_5\}$, then both $(s, i, t) \in E_i$ (since $s_{X_i} = (3, 1) = t_{X_i}$) and $(s, j, t) \in E_j$ (since $s_{X_j} = (1, 2) = t_{X_j}$). Fourthly, the description of any *walk* within the graph has the following generic form

$$(t^0, i(1), t^1, i(2), t^2, \dots, t^{k-1}, i(k), t^k)$$

which records not only the points $t^0, t^1, \dots, t^k$ touched on by the walk, but also the adjacencies $E_{i(1)}, \dots, E_{i(k)}$ used in passing from point to point. For example, assuming $X_i = \{T_2, T_3\}$, $X_j = \{T_3, T_5\}$, and $X_h = \{T_1, T_4, T_5\}$, the following expressions describe two different walks with the same initial and terminal points:

$$((2, 1, 3, 2, 3), j, (2, 2, 3, 1, 3), h, (2, 3, 1, 1, 3)),$$
$$((2, 1, 3, 2, 3), h, (2, 3, 1, 2, 3), i, (1, 3, 1, 2, 3), j, (2, 3, 1, 1, 3)).$$

Lastly, the assumption (22) ensures that the graph is *connected*. Indeed, the fact that the conditioned variables $Y_1, \dots, Y_m$ (on which any difference in coordinates is permitted) exhaust the full variable T allows us to transform any point s into any other point t through a sequence of changes each of which preserves adjacency.

The graphical structure described thus far is only determined by the set $\{X_1, \dots, X_m\}$ of the conditioning variables in the assumed set of kernels $\{p_1(Y_1|X_1), \dots, p_m(Y_m|X_m)\}$. As families of density functions, these kernels allow us to endow that structure with a *valuation function*. Specifically, let us consider any kernel $p_i(Y_i|X_i)$ in the set, the corresponding adjacency E_i (determined by X_i according to (25)), and any line (s, i, t) belonging to E_i (so that the projections s_{X_i} and t_{X_i} are equal and are a point in X_i°, whereas the projections s_{Y_i} and t_{Y_i} are possibly different points in Y_i°). The kernel $p_i(Y_i|X_i)$ provides definite values $p_i(s_{Y_i}|s_{X_i})$ and $p_i(t_{Y_i}|t_{X_i})$, which under the positivity condition (24) are both positive real numbers. Thus, they may be combined by division to obtain a positive value associated with the line in question:

$$R(s, i, t) = \frac{p_i(t_{Y_i}|t_{X_i})}{p_i(s_{Y_i}|s_{X_i})}. \tag{26}$$

This value may be interpreted as an odds quantity, being the ratio between the probabilities of two events $Y_i = t_{Y_i}$ and $Y_i = s_{Y_i}$ concerning the variable Y_i, both conditional on the event $X_i = t_{X_i} = s_{X_i}$ concerning the variable X_i. By applying this method in relation to each kernel $p_i(Y_i|X_i)$ in the given set and each line (s, i, t) in the corresponding adjacency E_i, a positive-valued function R on the pooled adjacency E is generated that upgrades the graphical structure in the following form:

$$(T^\circ, E, R) = (T^\circ, E_1 \cup \cdots \cup E_m, R).$$

In graph-theoretic terms, this is a line-valued directed multi-graph (Yao, Chen and Wang, 2014, p. 2).

The valuation R, which is first defined on single lines in the graph, can be consistently extended to any walk by setting

$$R(t^0, i(1), t^1, i(2), t^2, \dots, t^{k-1}, i(k), t^k) = \\ R(t^0, i(1), t^1) \cdot R(t^1, i(2), t^2) \cdots R(t^{k-1}, i(k), t^k). \tag{27}$$

Under condition (24), all factors in this product exist as positive real numbers, meaning that the product itself is a positive real number. The characterization of compatibility expressed in the next theorem specifically refers to the values that may result from this multiplicative formula (a product of odds).

Theorem 3. *Let $\{p_1(Y_1|X_1), \dots, p_m(Y_m|X_m)\}$ be a set of kernels that satisfy the conditions of saturation* (21), *exhaustiveness* (22), *and positivity* (24), *and let (T°, E, R) be the line-valued directed graph that can be constructed based on the kernels in the way described above. The kernels are mutually compatible if and only if the valuation R assigns the value* 1 *to every closed walk in the graph.*

Proof. "Only if" part. Suppose that the saturated kernels $\{p_1(Y_1|X_1), \dots, p_m(Y_m|X_m)\}$ are mutually compatible, that is, there exists a full density $r(T)$ such that $p_i(Y_i|X_i) = r(Y_i|X_i)$ for all $i = 1, \dots, m$, which means

$$p_i(t_{Y_i}|t_{X_i}) = r(t_{Y_i}|t_{X_i}) = \frac{r(t)}{r(t_{X_i})} \text{ for all } t \in T^\circ. \tag{28}$$

If $(t^0, i(1), t^1, i(2), t^2, \dots, t^{k-1}, i(k), t^0)$ is any closed walk in the graph (T°, E, R) (note the term t^0 in the place of t^k), then

$$\begin{aligned} R(t^0, i(1), t^1, \dots, t^{k-1}, i(k), t^0) &= R(t^0, i(1), t^1) \cdots R(t^{k-1}, i(k), t^0) \\ &= \frac{p_{i(1)}(t^1_{Y_{i(1)}}|t^1_{X_{i(1)}})}{p_{i(1)}(t^0_{Y_{i(1)}}|t^0_{X_{i(1)}})} \cdots \frac{p_{i(k)}(t^0_{Y_{i(k)}}|t^0_{X_{i(k)}})}{p_{i(k)}(t^{k-1}_{Y_{i(k)}}|t^{k-1}_{X_{i(k)}})} \\ &= \frac{r(t^1)/r(t^1_{X_{i(1)}})}{r(t^0)/r(t^0_{X_{i(1)}})} \cdots \frac{r(t^0)/r(t^0_{X_{i(k)}})}{r(t^{k-1})/r(t^{k-1}_{X_{i(k)}})} \\ &= \frac{r(t^1)}{r(t^0)} \cdots \frac{r(t^0)}{r(t^{k-1})} = \frac{r(t^1) \cdots r(t^{k-1}) \cdot r(t^0)}{r(t^0) \cdot r(t^1) \cdots r(t^{k-1})} = 1 \end{aligned}$$

where the first three equalities are justified by (27), (26), and (28), respectively, and the fourth is derived from the fact that $(t^h, i(h+1), t^{h+1}) \in E_{i(h+1)}$ for every $h = 0, \dots, k-1$, meaning that $t^h_{X_{i(h+1)}} = t^{h+1}_{X_{i(h+1)}}$.

"If" part. Let $\{p_1(Y_1|X_1), \dots, p_m(Y_m|X_m)\}$ be a set of saturated kernels such that in the resulting graph (T°, E, R) the valuation R assigns value 1 to every closed walk. First note that definition (26) implies $R(s, i, t) = 1/R(t, i, s)$ for each pair $\{(s, i, t), (t, i, s)\}$ of symmetric lines. In turn, this implies that our assumption concerning R can be reformulated as follows:

for all walks $(t^0, i(1), t^1, \dots, t^{k-1}, i(k), t^k)$ and $(u^0, j(1), u^1, \dots, u^{h-1}, j(h), u^h)$ (29)
if $t^0 = u^0$ and $t^k = u^h$
then $R(t^0, i(1), t^1, \dots, t^{k-1}, i(k), t^k) = R(u^0, j(1), u^1, \dots, u^{h-1}, j(h), u^h)$.

Consider the following constructive procedure: (arbitrarily) choose a reference point o in the space T°; construct a function f on T° by setting

$$f(t) \text{ equal to the value assigned by } R \text{ to any walk from } o \text{ to } t, \text{ for all } t \in T^\circ; \tag{30}$$

and then define

$$r(t) = \frac{f(t)}{\sum_{s\in T^\circ} f(s)} \text{ for all } t \in T^\circ. \tag{31}$$

The connectedness of the graph (ensured by (22)), the positivity assumption (24), and the hypothesis (29) imply that f is a well-defined positive-valued function over the space T° (in particular, hypothesis (29) ensures that for each $t \in T^\circ$, the value $f(t)$ is invariant relative to the available walks from o to t). Hence, the function $r(T)$ specified by (31) is a well-defined full density over the space T°. We will now show that $r(T)$ is indeed a consensus density for the given kernels, that is

$$p_i(t_{Y_i}|t_{X_i}) = r(t_{Y_i}|t_{X_i}) = \frac{r(t)}{r(t_{X_i})} \text{ for all } t \in T^\circ \text{ and all } i = 1, \ldots, m \tag{32}$$

which allows us to conclude that the given kernels are mutually compatible. Consider any $i = 1, \ldots, m$ and any $t \in T^\circ$, and define

$$T^\circ|(i,t) = \{s \in T^\circ : (t,i,s) \in E_i\} = \{s \in T^\circ : t_{X_i} = s_{X_i}\}$$

so that

$$\sum_{s\in T^\circ|(i,t)} p_i(s_{Y_i}|s_{X_i}) = 1 \quad \text{and} \quad \sum_{s\in T^\circ|(i,t)} r(s) = r(t_{X_i}).$$

If $W = (o, i(1), w^1, \ldots, w^{k-1}, i(k), t)$ is any walk from o to t, then $f(t) = R(W)$, and for each $s \in T^\circ|(i,t)$ the list $(o, i(1), w^1, \ldots, w^{k-1}, i(k), t, i, s)$ describes a walk from o to s, so that from (26), (27), and (30)

$$f(s) = R(W) \cdot \frac{p_i(s_{Y_i}|s_{X_i})}{p_i(t_{Y_i}|t_{X_i})}.$$

Therefore

$$\frac{r(t)}{r(t_{X_i})} = \frac{r(t)}{\sum\limits_{s\in T^\circ|(i,t)} r(s)} = \frac{f(t)}{\sum\limits_{s\in T^\circ|(i,t)} f(s)} = \frac{R(W)}{\sum\limits_{s\in T^\circ|(i,t)} R(W) \cdot \frac{p_i(s_{Y_i}|s_{X_i})}{p_i(t_{Y_i}|t_{X_i})}} = \frac{p_i(t_{Y_i}|t_{X_i})}{\sum\limits_{s\in T^\circ|(i,t)} p_i(s_{Y_i}|s_{X_i})} = p_i(t_{Y_i}|t_{X_i}),$$

which verifies Equation (32). □

A key idea underlying the proof of Theorem 3 is that the *product* of a suitable *chain of ratios* between *conditional* probabilities (that is, a product of odds) equals the *ratio* between two *joint* probabilities associated with the first and the last links in the chain. The paper generally cited as the source of this idea is Besag (1974), where this odds-product method is applied to problems of spatial statistics. The same idea occurs as a crucial principle in several studies concerning compatibility of distributions, although from one study to another there may be differences in the mathematical context in which it is embedded and the form in which it is expressed (Gurevich, 1992, pp. 373–374; Cressie, 1993, pp. 412–414; Hobert and Casella, 1998, pp. 48–49; Slavkovic and Sullivant, 2006, pp. 198 and 206; Yao, Chen and Wang, 2014). In particular, Slavkovic and Sullivant (2006) address the compatibility problem using tools from the algebra of polynomials, show that the compatibility between any two or more saturated probability kernels can be characterized in terms of a system of binomial equations, and come to conclude that for any fixed combination $\{(Y_1|X_1), \ldots, (Y_m|X_m)\}$ of saturated variable pairs, the set of all combinations $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ of kernels that are mutually compatible (within each combination) is a "unimodular toric variety". The algebraic method applied in the cited study is different from the graph-theoretic view we took in preparing and proving Theorem 3 above. Nevertheless, there is a common root in both approaches, which may be recognized by comparing the "circuits" used in constructing the required binomial equations with the "closed walks" mentioned in Theorem 3, and by considering this simple fact: if $(v_1, v_2, v_3, v_4, \ldots, v_{r-1}, v_r)$ is a list of an even number of non-null values ("indeterminates" in polynomials), then the binomial equation $v_1 v_3 \cdots v_{r-1} - v_2 v_4 \cdots v_r = 0$ is equivalent to the equation $(v_1/v_2)(v_3/v_4) \cdots (v_{r-1}/v_r) = 1$ concerning a product of ratios. Ultimately, that common root is related to the Besag's (1974) idea of a consistent chain of pairs of conditional probabilities mentioned above. The Slavkovic and Sullivant's (2006) method has the additional merit of wider generality: it can also be applied to saturated kernels that (while satisfying the necessary condition (23)) may violate the condition (24) of positivity.

One advantage of the characterization in Theorem 3 (also shared by that in Slavkovic and Sullivant, 2006) is that it is of deductive type: the mutual compatibility of the given kernels $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ can be decided through a series of tests on the graphical structure (T°, E, R), which is directly deducible from the kernels in question. A limitation is that it is applicable only to saturated kernels, or more generally to kernels in which the united variable $Y_i \cup X_i$ is the same for all $i = 1, \ldots, m$. Another limitation is its expensiveness, as the acceptance of the compatibility hypothesis requires a test (with positive result) on each cycle within the graph (T°, E, R). Algorithms may be devised to simplify this testing process by exploiting redundancies implicit in the graph (Wang and Kuo, 2010; Kuo and Wang, 2011; Yao, Chen and Wang, 2014). Lastly we remark that, of the cases covered by Theorem 3, special notice should be given to the case in which the conditioned variables $Y_1, \ldots, Y_m$ in the kernels, in addition to being exhaustive, are also mutually disjoint, so that they form a partition of the full variable T – this is the "alternating scheme" focused on by Lemma 3(ii). In this case, a formal simplification is available, since if $Y_i \cap Y_j = \varnothing$, $X_i = T \setminus Y_i$, and $X_j = T \setminus Y_j$, then for no points $s \neq t$ in T° can we have both $s_{X_i} = t_{X_i}$ and $s_{X_j} = t_{X_j}$; hence, the directed graph (T°, E) is simple, that is, a graph with at most one line directed from a point s to another point t.

The proof of the "if" part of Theorem 3 implies that, under the stated conditions, the given kernels admit a single consensus density. Indeed, under those conditions, the graph (T°, E, R) is connected, so that the rule in (30) determines a single valuation $f(T)$ over T°, which in turn by (31) determines a single full density $r(T)$ dominating over the given kernels (this argument, if applied to an alternating scheme of kernels, provides a supplementary proof of Lemma 3(ii)). The connectedness of the graph and the positive valuation of all lines in it are ensured by assumptions (22) and (24). Now suppose that the kernels $\{p_1(Y_1|X_1), \ldots, p_m(Y_m|X_m)\}$ satisfy (22) (as well as (23)) but violate (24), so that there are points u in

T° such that $p_i(u_{Y_i}|u_{X_i}) = 0$ for some $i = 1, \dots, m$. In this looser situation, the pooled adjacency E will contain lines (s, i, t) such that the ratio $R(s, i, t)$ as defined by (26) is either null or nonexistent as a real number; hence, these lines are of no use for the odds-product method and should be cancelled from E, which then degrades into a poorer adjacency G. Unlike E, this G could fail to be connected, so that the space T° would be divided into two or more connected components $(T^\circ)_1, \dots, (T^\circ)_k$ according to G. If the graph satisfies the condition stated in Theorem 3 (concerning closed walks), then by the odds-product method we would still be able to uniquely construct densities $r_1(T), \dots, r_k(T)$ separately defined on $(T^\circ)_1, \dots, (T^\circ)_k$. These could be assembled into a full density $r(T)$ by choosing a set $(c_1, \dots, c_k)$ of positive numbers with unit sum and then setting

$$r(t) = c_1 r_1'(t) + \cdots + c_k r_k'(t) \qquad \text{for all } t \in T^\circ, \tag{33}$$

where (for $h = 1, \dots, k$) we set $r_h'(t) = r_h(t)$ if $t \in (T^\circ)_h$ and $r_h'(t) = 0$ otherwise. Following the reasoning developed for Theorem 3 and considering that any two $(T^\circ)_h \neq (T^\circ)_{h'}$ have disjoint projections on each space X_i° (for $i = 1, \dots, m$), it can be seen that a density $r(T)$ thus constructed dominates each of the given kernels, so that these are mutually compatible. The main point of this discussion is that if the diminished graph (T°, G) fails to be connected, then there are several (infinitely many) consensus densities that are constructible according to (33), simply because the set $(c_1, \dots, c_k)$ may be chosen as any k-tuple of positive numbers with unit sum. Thus, the possible violation of the positivity condition (24) is detrimental not to the existence of a consensus density (which is still guaranteed by the condition on closed walks stated in Theorem 3), but to the uniqueness of such a density.

6. Concluding remarks: the varying saliency of the compatibility problem

In principle, as noted in the Introduction, compatibility of distributions is a basic requirement of any probabilistic model. Indeed, if the distributional assumptions in a model were not fully compatible with one another, then the data analyses guided by the model could in fact be directed towards a nonexistent ideal target, as there could be no global distribution that consistently encompasses all the local distributions postulated by the assumptions. In practice, however, the compatibility problem does not appear to have equal saliency for different kinds of probabilistic models. There are even models for which that problem may seem an idle question, since compatibility appears implicit in the basic structure of such models, regardless of the mathematical form of the distributions involved. In this concluding section, we will use some of the results of our study (in particular, those in Section 3) to illustrate the reasons for the different saliency of compatibility relative to different elementary kinds of probabilistic models.

First, we consider a simple model of classical statistics: the model for comparing the means of two normal populations under the assumption of equal variance. The full variable in the model is the set $T = \{T_{1,1}, \dots, T_{1,n_1}, T_{2,1}, \dots, T_{2,n_2}\}$, formed of a sample taken from one population and a sample taken from the other. In addition to the assumptions of stochastic independence within and between both samples, the model includes distributional assumptions, which are expressed by these assignments

$$T_{1,i} \sim \text{Normal}(\mu_1, \sigma^2) \text{ and } T_{2,j} \sim \text{Normal}(\mu_2, \sigma^2) \qquad \text{for } i = 1, \dots, n_1 \text{ and } j = 1, \dots, n_2, \tag{34}$$

where μ_1, μ_2, and σ^2 are quantities that are unknown but (in the classical view) not treated as random variables. In the terms used in this study, the distributional assumptions are about the following marginal densities

$$p(T_{1,1}), \ldots, p(T_{1,n_1}), p(T_{2,1}), \ldots, p(T_{2,n_2})$$

which may be viewed as probability kernels on the variable pairs

$$(T_{1,1}|\varnothing), \ldots, (T_{1,n_1}|\varnothing), (T_{2,1}|\varnothing), \ldots, (T_{2,n_2}|\varnothing)$$

which in turn are atoms in the lattice $\widetilde{\mathcal{O}}(T)$. The (elementary) conditioned variables in the pairs are disjoint from one another, and the incidence relation $\to$ when referred to these pairs is acyclic (indeed, it is empty), so that from Theorem 1, the $n_1 + n_2$ kernels (or marginal densities) are compatible with one another for any values of μ_1, μ_2, and σ^2. Thus, distributional compatibility is assured here by the basic structure of the model (regardless of the form of the distributions) and this may explain why the compatibility question is not generally raised when presenting this and other similar models of classical statistics. Of course, there is another, more familiar way of reaching the same conclusion, that is, to observe that the product function $p(T) = p(T_{1,1}) \cdots p(T_{1,n_1}) \cdot p(T_{2,1}) \cdots p(T_{2,n_2})$ is certainly a consensus density for the $n_1 + n_2$ marginal densities, as each of these can be deduced from $p(T)$ by projection (operation J in Definition 3). Note that if the assumptions of stochastic independence are left out of the model, then the product function is not the only consensus density for the given marginal densities.

Our second example is a Bayesian expansion of the preceding classical model. Specifically, let us suppose that the parameters μ_1, μ_2, and σ^2 are themselves conceived of as random variables and that the system (34) becomes enriched by the following distributional assumptions:

$$\begin{gathered}\mu_1 \sim \text{Normal}(\nu, \tau^2), \qquad \mu_2 \sim \text{Normal}(\nu, \tau^2), \qquad \sigma^2 \sim \text{InverseGamma}(\alpha, \beta), \\ \nu \sim \text{Uniform}(0, 50), \qquad \tau^2 \sim \text{Uniform}(0, 10), \qquad \alpha \sim \text{Uniform}(0, 1), \qquad \beta \sim \text{Uniform}(0, 1).\end{gathered}$$

The full set of elementary random variables in the model then becomes

$$T = \{T_{1,1}, \ldots, T_{1,n_1}, T_{2,1}, \ldots, T_{2,n_2}, \mu_1, \mu_2, \sigma^2, \nu, \tau^2, \alpha, \beta\}$$

in which the first $n_1 + n_2$ elements are the data, the next three are first-level parameters, and the last four are second-level parameters (or hyper-parameters). On the whole, the distributional assumptions in the model are constraints on these probability kernels

$$\begin{gathered}p(T_{1,1}|\mu_1, \sigma^2), \ldots, p(T_{1,n_1}|\mu_1, \sigma^2), \; p(T_{2,1}|\mu_2, \sigma^2), \ldots, p(T_{2,n_2}|\mu_2, \sigma^2), \\ p(\mu_1|\nu, \tau^2), \; p(\mu_2|\nu, \tau^2), \; p(\sigma^2|\alpha, \beta), \; p(\nu|\varnothing), \; p(\tau^2|\varnothing), \; p(\alpha|\varnothing), \; p(\beta|\varnothing).\end{gathered}$$

Note that the assumptions precisely specify the densities of the four hyper-parameters, and thus the conditioning variable in the last four kernels is the empty variable. Figure 3 is the directed graph generated by the incidence relation $\to$ when this is applied to the set of variable pairs in the kernels in question. It is seen that the graph has no cycle. This property and the fact that the conditioned variables in the pairs are mutually disjoint (they are distinct elementary variables) guarantee (again by Theorem 1) that this set

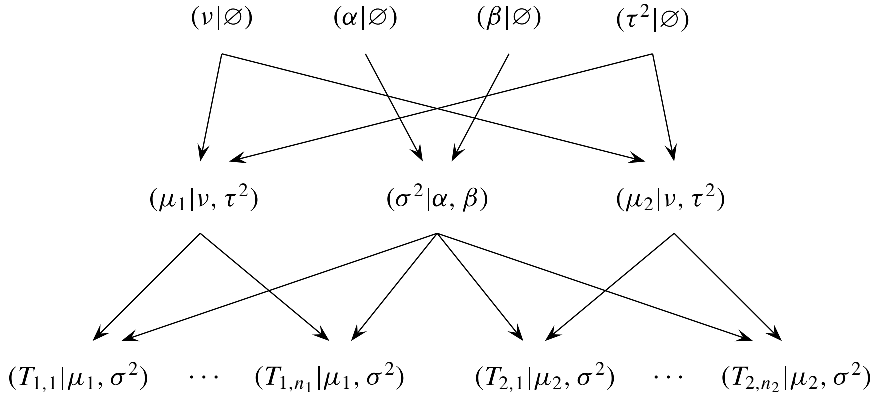


Figure 3. Incidence relation within a set of $n_1 + n_2 + 7$ variable pairs.

of variable pairs has sure compatibility, so that the distributional assumptions in the model are compatible. This appears to be a general property of hierarchical Bayesian models (Gelman et al., 2014, chapter 5; Lunn et al., 2013, chapter 10): the hierarchical form of a model implies absence of cycles among variable pairs in the primitive kernels, and then mutual compatibility of the kernels themselves.

Similar arguments may be used for the probabilistic models known as "Bayesian networks" (Pearl, 1988). Any Bayesian network rests on a graphical structure called a directed acyclic graph (DAG). A DAG differs from the kind of structures illustrated in Figure 3, since the nodes in it are individual elementary variables, rather than pairs of variables. However, a DAG can be faithfully translated into a graph of variable pairs, simply by replacing each elementary variable T_i by the pair $(T_i|X_i)$, where X_i is the set of "parents" of T_i (i.e., variables sending an arrow towards T_i in the DAG). For example, using this criterion, the graph of individual variables in the left-hand part of Figure 4 becomes translated into the graph of variable pairs in the right-hand part of the same figure. A DAG is acyclic, and this implies that the corresponding graph of variable pairs (related by the incidence relation $\rightarrow$) is also acyclic. Hence, by virtue of Theorem 1, we have that any assignment of kernels $\{p_1(T_1|X_1), \ldots, p_n(T_n|X_n)\}$ (which specify how each variable T_i is expected to *depend* on the set X_i of its parents in the network) will be mathematically consistent, that is, there is a consensus density $p(T)$ for the kernels. Within this structural assurance of consistency presumably lies a reason for the importance of the acyclicity requirement for Bayesian networks. Moreover, the DAG of a Bayesian network is also intended to represent a set of conditional stochastic *independencies* between the variables in the network. Specifically, each T_i is assumed to be independent of $T \setminus (T_i \cup X_i \cup Z_i)$ conditional on X_i, where X_i is the set of parents and Z_i is the set of descendants of T_i in the DAG. For example, the DAG in Figure 4 is intended to represent the following conditional independencies:

$$I(T_1, T_2|\varnothing),\ I(T_2, T_1T_3|\varnothing),\ I(T_3, T_2T_4|T_1),\ I(T_4, T_3|T_1T_2),\ I(T_5, T_1T_2|T_3T_4).$$

Based on these, any set of hypothesized kernels

$$p(T_1|\varnothing),\ p(T_2|\varnothing),\ p(T_3|T_1),\ p(T_4|T_1T_2),\ p(T_5|T_3T_4)$$

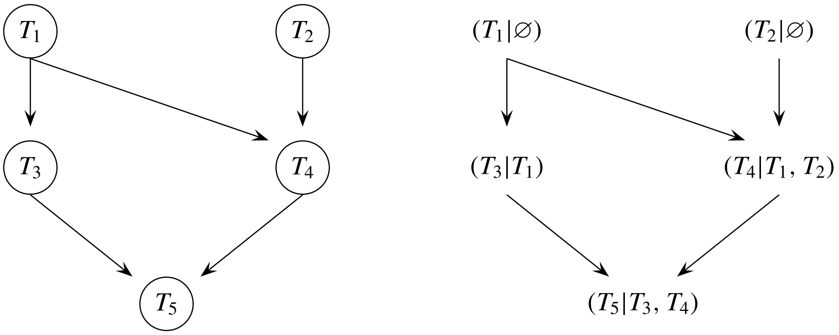


Figure 4. The DAG of a Bayesian network for five variables (left) and its rewriting in terms of incidence between variable pairs (right).

may equivalently be presented as

$$p(T_1|\varnothing),\ p(T_2|T_1),\ p(T_3|T_1T_2),\ p(T_4|T_1T_2T_3),\ p(T_5|T_1T_2T_3T_4).$$

In this form, the kernels constitute a cumulative scheme, so that by Lemma 3(i) there is a single full density $p(T)$ that is admissible for the network, which is obtainable from the kernels through multiple promotion (or the "chain rule" for Bayesian networks: Pearl, 1988, pp. 119–120; Kjærulff and Madsen, 2008, pp. 58–60).

As for Bayesian networks, the definition of a Markov random field is a mix of stochastic independence assumptions (each elementary variable T_i is assumed to be stochastically independent of $T \setminus (T_i \cup X_i)$ conditional on X_i, this being the set of elementary variables adjacent to T_i in the field) and stochastic dependence assumptions (a kernel $p(T_i|X_i)$ or a class of such kernels is associated with each elementary variable T_i and expresses how T_i is expected to be affected by its neighborhood X_i) (Kindermann and Snell, 1980; Koller and Friedman, 2009, chapter 4). Unlike the DAG of a Bayesian network, however, the graphical structure (an undirected graph) implicit in a Markov field does not generally possess the acyclicity character required for directly ensuring (by Theorem 1) compatibility between the kernels. Let us consider, for example, the graph in Figure 5. To specify a Markov field on this graph is tantamount to specifying nine local kernels

$$\begin{aligned}
&p(T_1|T_2T_4),\ p(T_2|T_1T_3T_5),\ p(T_3|T_2T_6),\\
&p(T_4|T_1T_5T_7),\ p(T_5|T_2T_4T_6T_8),\ p(T_6|T_3T_5T_9),\\
&p(T_7|T_4T_8),\ p(T_8|T_5T_7T_9),\ p(T_9|T_6T_8).
\end{aligned}$$

Each of these expresses how one of the nine elementary variables is assumed to be affected by its neighborhood in the graph. Evidently, there are $\rightarrow$-cycles within the set of variable pairs – for example, $(T_1|T_2T_4) \rightarrow (T_2|T_1T_3T_5) \rightarrow (T_1|T_2T_4)$ is a cycle – so that Theorem 1 cannot be invoked to directly conclude in favor of compatibility between the kernels. To answer the compatibility question in this situation, we need to consider the numerical properties of the kernels themselves, as families of density functions. The theory of Markov random fields especially explores the stochastic independence properties

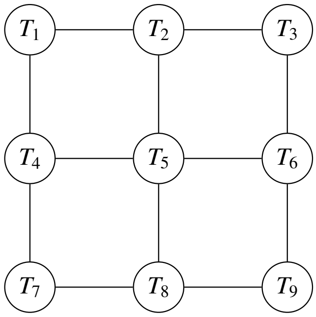


Figure 5. The neighborhood system of a possible Markov random field on nine variables.

implicit in such fields. For the reasons now suggested, however, advancements regarding compatibility, beyond structural assurance and in the directions illustrated in Sections 4 and 5, may also be relevant in dealing with such probabilistic models (Kaiser and Cressie, 2000; Kaiser, 2002).

Our last comment is on the relevance of the compatibility requirement for the so-called "Gibbs sampler" in probability simulations (Geman and Geman, 1984, pp. 730–732; Robert and Casella, 2004, chapter 10). A typical context for a Gibbs sampling is formed of a set $T = \{T_1, \ldots, T_n\}$ of elementary variables and a corresponding complete set $\{p(T_1|T \setminus \{T_1\}), \ldots, p(T_n|T \setminus \{T_n\})\}$ of elementary saturated kernels (so-called "full conditionals"). The method produces simulations of the full density $p(T)$ implied by the n input kernels, as well as simulations of other densities or probability kernels dominated by $p(T)$. In most applications, the input kernels are specified in a deductive manner. This means that a researcher first specifies the analytical expression for a full density $p(T)$, and then deduces (by conditioning) the analytical expression for each input kernel $p(T_i|T \setminus \{T_i\})$, which should be "available to sampling", that is, for each $x_i \in (T \setminus \{T_i\})^\circ$, it is practically possible to simulate samples from the distribution $p(T_i|x_i)$. In this approach, the mutual compatibility of the input kernels is true by construction, simply because their analytical expressions are deduced from the formula of $p(T)$, so that the input kernels are jointly dominated by $p(T)$. Furthermore, the kernels $\{p(T_1|T \setminus \{T_1\}), \ldots, p(T_n|T \setminus \{T_n\})\}$ thus determined form an alternating scheme, so that under the conditions stated in Lemma 3(ii) they constitute a sufficient basis for the univocal recovery of $p(T)$. A different route, however, could be taken. A researcher could directly specify (rather than deduce from a full density $p(T)$) a set $\{p(T_1|T \setminus \{T_1\}), \ldots, p(T_n|T \setminus \{T_n\})\}$ of elementary saturated kernels, and then apply to these the Gibbs sampler for simulation purposes. It is from this perspective that the critical role of the compatibility requirement appears more clearly. The mutual compatibility of the input kernels thus proposed has neither deductive assurance (they are not deduced from a common parent density $p(T)$) nor structural assurance (the incidence → among their variable pairs forms a complete directed graph, certainly containing cycles, so that the "if" part of Theorem 1 cannot be invoked). Thus, the proposed input kernels may fail to be compatible, in which case applying the Gibbs sampler would be tantamount to trying to simulate a nonexistent full density, resulting in erratic non-converging output sequences (Heckerman et al., 2000, pp. 56–57; Robert and Casella, 2011, p. 108). All of this demonstrates the saliency of the compatibility problem for the Gibbs sampler in current use for probability simulations.

References

Arnold, B. C., Castillo, E. and Sarabia, J. M. (1999), *Conditional specification of statistical models*. New York: Springer.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2001), “Conditionally specified distributions: An introduction (with comments and a rejoinder by the authors)”, *Stat. Science* **16**, 249–274. DOI 10.1214/ss/1009213728

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2002), “Exact and near compatibility of discrete conditional distributions”, *Comput. Stat. Data Anal.* **40**, 231–252. DOI 10.1016/S0167-9473(01)00111-6

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2004), “Compatibility of partial or complete conditional probability specifications”, *J. Stat. Planning and Inference* **123**, 133–159. DOI 10.1016/S0378-3758(03)00137-X

Arnold, B. C. and Press, S. J. (1989), “Compatible conditional distributions”, *J. Amer. Stat. Assoc.* **84**, 152–156. DOI 10.2307/2289858

Berti, P., Dreassi, E. and Rigo, P. (2014), “Compatibility results for conditional distributions”, *J. Multivariate Anal.* **125**, 190–203. DOI 10.1016/j.jmva.2013.12.009

Besag, J. E. (1974), “Spatial interaction and the statistical analysis of lattice systems (with discussion)”, *J. Royal Stat. Society, Series B* **36**, 192–236. DOI 10.1111/j.2517-6161.1974.tb00999.x

Billingsley, P. (1995), *Probability and measure*. New York: Wiley.

Burigana, L. and Vicovaro, M. (2020), “Inferring properties of probability kernels from the pairs of variables they involve”, *Algebraic Stat.* **11**, 79–97. DOI 10.2140/astat.2020.11.79

Chang, J. T. and Pollard, D. (1997), “Conditioning as disintegration”, *Stat. Neerlandica* **51**, 287–317. DOI 10.1111/1467-9574.00056

Chen, H. Y. (2010), “Compatibility of conditionally specified models”, *Stat. Prob. Let.* **80**, 670–677. DOI 10.1016/j.spl.2009.12.025

Cressie, N. (1993), *Statistics for spatial data*. New York: Wiley.

Gelfand, A. E. and Smith, A. F. M. (1990), “Sampling-based approaches to calculating marginal densities”, *J. Amer. Stat. Assoc.* **85**, 398–409. DOI 10.2307/2289776

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014), *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A. and Speed, T. P. (1993), “Characterizing a joint probability distribution by conditionals”, *J. Royal Stat. Society, Series B* **55**, 185–188. DOI 10.1111/j.2517-6161.1993.tb01477.x

Geman, S. and Geman, D. (1984), “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*, 721–741. DOI 10.1109/TPAMI.1984.4767596

Gourieroux, C. and Monfort, A. (1979), “On the characterization of a joint probability distribution by conditional distributions”, *J. Econometrics* **10**, 115–118. DOI 10.1016/0304-4076(79)90070-8

Griffeath, D. (1976), "Introduction to random fields", pp. 425–458 in *Denumerable Markov chains*, edited by J. G. Kemeny, J. L. Snell and A. W. Knapp. Berlin: Springer.

Gurevich, B. M. (1992), "On the joint distribution of random variables with given cross conditional distributions: discrete case", *Theory of Probability and its Applications* **36**, 371–375. DOI 10.1137/1136041

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. and Kadie, C. (2000), "Dependency networks for inference, collaborative filtering, and data visualization", *J. Machine Learning Research* **1**, 49–75.

Hobert, J. P. and Casella, G. (1998), "Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors", *J. Comput. Graphical Stat.* **7**, 42–60. DOI 10.1080/10618600.1998.10474760

Kaiser, M. S. (2002), "Markov random field models", pp. 1213–1224 in *Encyclopedia of Environmetrics*, vol. 3, edited by A. H. El-Shaarawi and W. W. Piegorsch, New York: Wiley. 10.1002/9781118445112.stat07479

Kaiser, M. S. and Cressie, N. (2000), "The construction of multivariate distributions from Markov random fields", *J. Multivariate Anal.* **73**, 199–220. DOI 10.1006/jmva.1999.1878

Kindermann, R. and Snell, J. L. (1980), *Markov random fields and their applications*. Providence, RI: American Mathematical Society.

Kjærulff, U. B. and Madsen, A. L. (2008), *Bayesian networks and influence diagrams: A guide to construction and analysis*. New York: Springer.

Koller, D. and Friedman, N. (2009), *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Koski, T. and Noble, J. M. (2009), *Bayesian networks: An introduction*. Chichester, UK: Wiley.

Kuo, K. L, Song, C. C. and Jiang, T. J. (2017), "Exactly and almost compatible joint distributions for high-dimensional discrete conditional distributions", *J. Multivariate Anal.* **157**, 115–123. DOI 10.1016/j.jmva.2017.03.005

Kuo, K. L. and Wang, Y. J. (2011), "A simple algorithm for checking compatibility among discrete conditional distributions", *Comput. Stat. Data Anal.* **55**, 2457–2462. DOI 10.1016/j.csda.2011.02.017

Lauritzen, S. L. (1996), *Graphical models*. Oxford, UK: Oxford University Press.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013), *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press.

Parthasarathy, K. R. (2005), *Introduction to probability and measure*. New Delhi: Hindustan Book Agency.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pollard, D. (2002), *A user's guide to measure theoretic probability*. Cambridge, UK: Cambridge University Press.

Robert, C. P. and Casella, G. (2004), *Monte Carlo statistical methods*. New York: Springer.

Robert, C. P. and Casella, G. (2011), “A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data”, *Stat. Science* **26**, 102–115. DOI 10.1214/10-STS351

Slavkovic, A. B. and Sullivant, S. (2006), “The space of compatible full conditionals is a unimodular toric variety”, *J. Symbolic Computation* **41**, 196–209. DOI 10.1016/j.jsc.2005.04.006

Tian, G. L., Tan, M., Ng, K. W. and Tang, M. L. (2009), “A unified method for checking compatibility and uniqueness for finite discrete conditional distributions”, *Commun. Stat. Theory Methods* **28**, 115–129. DOI 10.1080/03610920802169586

Wang, Y. J. and Kuo, K. L. (2010), “Compatibility of discrete conditional distributions with structural zeros”, *J. Multivariate Anal.* **101**, 191–199. DOI 10.1016/j.jmva.2009.07.007

Yao, Y. C., Chen, S. C. and Wang, S. H. (2014), “On compatibility of discrete full conditional distributions: A graphical representation approach”, *J. Multivariate Analysis* **124**, 1–9. DOI 10.1016/j.jmva.2013.10.007

Received 2020-01-25. Revised 2020-06-11. Accepted 2020-07-06.

LUIGI BURIGANA: luigi.burigana@unipd.it
Department of General Psychology, University of Padua, I-35131 Padova, Italy

MICHELE VICOVARO: michele.vicovaro@unipd.it
Department of General Psychology, University of Padua, I-35131 Padova, Italy