11:1 2020

Algebraic Statistics

INFERRING PROPERTIES OF PROBABILITY KERNELS FROM THE PAIRS OF VARIABLES THEY INVOLVE

LUIGI BURIGANA AND MICHELE VICOVARO

msp

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.79

INFERRING PROPERTIES OF PROBABILITY KERNELS FROM THE PAIRS OF VARIABLES THEY INVOLVE

LUIGI BURIGANA AND MICHELE VICOVARO

A probabilistic model may involve families of probability functions such that the functions in a family act on a definite (possibly multiple) variable and are indexed by the values of some other (possibly multiple) variable. "Probability kernel" is the term here adopted for referring to any one such family. This study highlights general properties of probability kernels that may be inferred from set-theoretic characteristics of the pairs of variables on which the kernels are defined. In particular, it is shown that any complete set of such pairs of variables has the algebraic form of a lattice, which is then inherited by any complete set of compatible kernels defined on those pairs; that on pairs of variables a criterion may be applied for testing whether corresponding probability kernels are compatible with one another and may thus be the building blocks of a consistent probabilistic model; and that the order between pairs of variables within their lattice provides a general diagnostic about deducibility relations between probability kernels. These results especially relate to models that involve a number of random variables and several interrelated conditional distributions acting on them; for example, hierarchical Bayesian models and graphical models in statistics, Bayesian networks and Markov fields, and Bayesian models in the experimental sciences.

1. Introduction

A general way of expressing the (possible) probabilistic dependence of a random variable Y on another random variable X is in terms of a family of probability distributions for Y that are conditional on the distinct possible values of X. If we denote by X° and Y° the spaces of the two variables — or, more concretely, the sets of their possible values — then such a family may formally be indicated as $(p(Y|x) : x \in X^\circ)$, where $p(Y|x)$ (for any $x \in X^\circ$) is the probability function on the domain Y° that would rule the variable Y under the condition $X = x$. If there are differences within the family — that is, if $p(Y|x) \neq p(Y|x')$ for some $x \neq x'$ belonging to X° — then there is some form of stochastic dependence of Y on X. Otherwise, the two variables are stochastically independent of each other. In this article, we refer to such a family of conditional probability functions by the name *probability kernel* and by the symbol $p(Y|X)$ — so that $p(Y|X)$ is the same as $(p(Y|x) : x \in X^\circ)$ by definition. The terms X and Y may be *multiple* variables and are here conceived as *disjoint* subsets of a *full* variable T, this being the collection of all elementary random variables involved in a probabilistic model. Typically, the probability functions constituting one kernel $p(Y|X)$ are of the same measure-theoretic type; for example, they may

Burigana is the corresponding author.
Keywords: probability kernel, conditional probability, compatibility, lattice, Bayesian model.

be either all mass functions ($p(Y|X)$ would be a kernel of discrete type) or all density functions ($p(Y|X)$ would be a kernel of continuous type)[1].

A probabilistic model with a suitably large set T of elementary variables may involve several such probability kernels $p(Y|X)$, $p(V|U)$, $p(Z|W)$, ... that concern (elementary or multiple) variables included in T. The two variables in any single kernel are assumed disjoint, but variables involved in different kernels may have non-empty intersections, which themselves count as variables included in T. The importance of any single kernel may be due either to its being a postulate in a given model, that is, a basic assumption characterizing the probabilistic dependence between relevant variables (or specifying the kind of distributions of those variables), or to its being a logical consequence of the postulates, which may be crucial for the interpretation of the model and its fitting to empirical data. Models involving several interrelated kernels may be generally referred to as *conditionally specified* probabilistic structures, for marking the central role played by assumptions of conditional distributions in the definition of such models (Arnold, Castillo, & Sarabia, 1999). Examples can be found in various areas of applied probability, such as graphical models in statistics (Koller & Friedman, 2009), hierarchical Bayesian modeling (Gelman et al., 2014, chapter 5), Bayesian networks in artificial intelligence (Darwiche, 2009), applications of Markov random fields (Blake, Kohli, & Rother, 2011), and Bayesian modeling in experimental sciences (Kersten, Mamassian, & Yuille, 2004; Rouder, Morey, & Pratte, 2017).

Any probability kernel $p(Y|X)$ has a definite *variable pair* $(Y|X)$ as its field of action, that is, an ordered pair formed of a conditioned variable Y (on the left of the bar) and a conditioning variable X (on the right of the bar) which are, in general, disjoint sub-variables of a full variable T. Variable pairs underlying distinct probability kernels may be subjected to comparisons, combinations, or transformations in mere set-theoretic terms, that is, as pairs of sets of elementary variables, irrespectively of the individual properties of the elementary variables they collect. Such set-theoretic manipulations may imply algebraic regularities worthy of note, and one may conjecture that these regularities concerning the variable pairs have meaningful consequences regarding the probability kernels acting on the variable pairs themselves. The aim of this article is precisely that of highlighting some general properties of probability kernels that may be inferred from the set-theoretic configuration of the variable pairs on which the kernels are defined.

In order to illustrate the association between actions on probability kernels and actions on the underlying variable pairs, let us refer to Figure 1, which represents a basic case in the theory of conditional probabilities. For simplicity, we suppose that X, W, and Z are elementary variables of discrete type, but the example is easily generalizable to multiple and/or continuous variables. Three basic operations on kernels are illustrated by the figure.

The first is *projection*. For example, one may pass from $p(W, Z|X)$ to $p(Z|X)$ by setting $p(z|x) = \sum_{w \in W^\circ} p(w, z|x)$ for all $(x, z) \in (X, Z)^\circ$ (this means: $x \in X^\circ$ and $z \in Z^\circ$).

[1] Our use of the word "kernel" is consistent with the meaning taken by this word in the theories of Markov processes and other probabilistic structures extensively involving conditional probability distributions (Meyn & Tweedie, 1993, p. 65; Lauritzen, 1996, p. 46). Indeed, the concept of a probability kernel, in such theories, generalizes the concept of a transition matrix in finite-state Markov chains, which amounts to an indexed set of conditional mass functions.

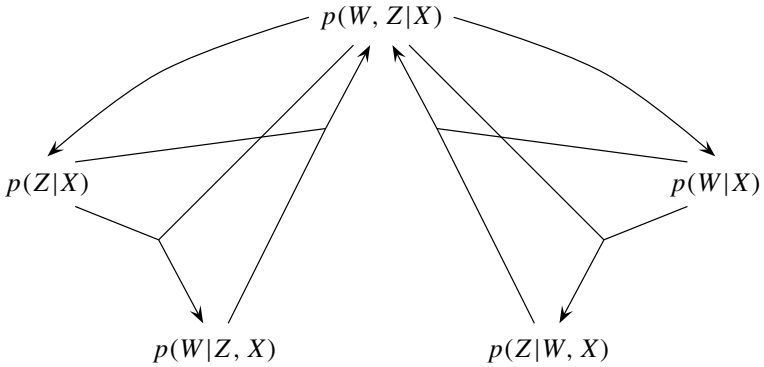

Figure 1. Probability kernels derivable from a top kernel $p(W, Z|X)$ through projection or conditioning.

The second is *conditioning*. For example, one may pass from $p(W, Z|X)$ to $p(W|Z, X)$ by setting $p(w|z, x) = p(w, z|x)/p(z|x)$ for all $(x, w, z) \in (X, W, Z)^\circ$, on presuming $p(z|x) > 0$ for all (x, z) in $(X, Z)^\circ$.

The third operation is *promotion*. For example, one may return from $p(W|Z, X)$ and $p(Z|X)$ to $p(W, Z|X)$ by setting $p(w, z|x) = p(w|z, x) \cdot p(z|x)$ for all $(x, w, z) \in (X, W, Z)^\circ$.

In this paper, the letters J, C, and M are used to denote proJection, Conditioning, and proMotion, respectively, so that the three moves just described may be symbolized as follows:

$$\begin{aligned} p(Z|X) &= J[p(W, Z|X), W] \\ p(W|Z, X) &= C[p(W, Z|X), Z] \\ p(W, Z|X) &= M[p(W|Z, X), p(Z|X)]. \end{aligned} \tag{1}$$

Of this example, what mostly matters for the aims of our study are the effects of the three operations on the variable pairs involved: projection J implies canceling a targeted component W from the left field (to the left of the bar), conditioning C implies moving a component Z from the left to the right field, and promotion M implies moving a component Z from the right to the left field of $p(W|Z, X)$ with the aid of a "promoter" $p(Z|X)$. We shall see that, based on these simple moves affecting the assignment of the variables to the left or the right fields in the probability kernels, algebraic constructs can be elaborated that have meaningful implications for the kernels at hand.

Our paper is formed of three main sections. Section 2 focuses on variable pairs and set-theoretic operations on them, and defines a binary relation that organizes any complete collection of such pairs as a lattice. In Section 3 the results obtained for variable pairs are transferred to probability kernels, and conditions are discussed that make it possible to ascend from kernels of low rank to kernels of higher rank in a lattice, which is a typical move in the construction of probabilistic models. In Section 4 the key binary relation between variable pairs will be shown to possess a general diagnostic ability about the deducibility relation between probability kernels.

2. Lattice of variable pairs

Let $T = \{T_1, \ldots, T_n\}$ be the complete set of elementary random quantities (observables, parameters, hyper-parameters, etc.) involved in a probabilistic model. By a *variable* we mean any subset of T. Thus, T itself is a variable, referred to as the *full variable* in the assumed model. Each singleton $\{T_i\}$ in T is an *elementary variable*. The symbol $\varnothing$ denotes the *empty variable*, which is the empty subset of T. Because variables are here understood as sets, statements such as "X is a sub-variable of Y" and "X and Y are disjoint variables", and formulas such as $X \subseteq Y$ and $X \cap Y = \varnothing$, are legitimate and meaningful in the language adopted in this paper.

A *variable pair* is any ordered pair $(Y|X)$ such that $X \cup Y \subseteq T$, $X \cap Y = \varnothing$, and $Y \neq \varnothing$, and the symbol $O(T)$ here denotes the complete collection of such pairs. Thus, if n is the cardinality of T, then $3^n - 2^n$ is the cardinality of $O(T)$. Besides, the symbol $\perp$ and the name *null variable pair* are here used for referring to any pair $(\varnothing|X)$ with $X \subseteq T$, and the symbol $\widetilde{O}(T)$ is a substitute for $O(T) \cup \{\perp\}$.

Drawing on equations (1), we define *projection* J, *conditioning* C, and *promotion* M on variable pairs by setting, for all $(Y|X) \in O(T)$:

$$J[(Y|X), W] = (Y \setminus W|X) \text{ for all } W \subset Y \tag{2}$$

$$C[(Y|X), W] = (Y \setminus W|W \cup X) \text{ for all } W \subset Y \tag{3}$$

$$M[(Y|X), (V|U)] = (Y \cup V|U) \text{ for all } (V|U) \in O(T) \text{ such that } V \cup U = X. \tag{4}$$

These definitions are designed so that reference to the null term $\perp$ is avoided. This limitation, however, can consistently be overcome by setting

$$\begin{aligned} &J[(Y|X), Y] = \perp, \quad C[(Y|X), Y] = \perp, \\ &M[\perp, (V|U)] = (V|U), \quad M[(Y|X), \perp] = (Y|X). \end{aligned} \tag{5}$$

Note, in particular, that the two additional equations concerning the M operation turn out to be consistent with rule (4) if $\perp$ becomes replaced by $(\varnothing|V \cup U)$ in the one equation and by $(\varnothing|X)$ in the other.

The equations in the next composite statement are self-evident:

$$\text{for all } (Y|X) \in O(T) \text{ and all } W, Z \subseteq Y \text{ such that } W \cap Z = \varnothing,$$

$$J[J[(Y|X), W], Z] = (Y \setminus (W \cup Z)|X) = J[J[(Y|X), Z], W] \tag{6}$$

$$C[C[(Y|X), W], Z] = (Y \setminus (W \cup Z)|W \cup Z \cup X) = C[C[(Y|X), Z], W] \tag{7}$$

$$J[C[(Y|X), W], Z] = (Y \setminus (W \cup Z)|W \cup X) = C[J[(Y|X), Z], W]. \tag{8}$$

They express invariance to change in order (commutativity) for combined J and C operations. In describing the result of the M operation, the rule is followed of writing the "to be promoted" variable pair $(Y|X)$ on the left, and its chosen "promoter" $(V|U)$ on the right, so that $X = V \cup U$. Therefore, if $M[(Y|X), (V|U)]$ is a syntactically correct formula, then $M[(V|U), (Y|X)]$ cannot be syntactically correct, because $U \neq Y \cup X$. For this simple reason, the M operation is not commutative. It is, however,

associative, because

$$\begin{aligned}&\text{for all } (Y|X), (V|U), (Z|W) \in O(T) \\ &\text{if } X = V \cup U \text{ and } U = Z \cup W, \\ &\text{then } M[M[(Y|X), (V|U)], (Z|W)] = (Y \cup V \cup Z|W) = M[(Y|X), M[(V|U), (Z|W)]].\end{aligned} \tag{9}$$

Furthermore, the following equations are easily proved:

$$C[M[(Y|X), (V|U)], V] = (Y|X) \tag{10}$$

$$J[M[(Y|X), (V|U)], Y] = (V|U). \tag{11}$$

These show how the operands of the M operation may be recovered from its result by suitably applying the C and J operations.

By combining the J and C operations, a binary relation between variable pairs is now defined, which is a key component of the theory in this study.

Definition 1. Let $(V|U)$ and $(Y|X)$ be variable pairs in $O(T)$. The former is *JC-derivable* from the latter (notation $(V|U) \preceq (Y|X)$) if $(V|U) = J[C[(Y|X), W], Z]$ for some W and Z disjoint sub-variables of Y. Furthermore, $\perp \preceq (Y|X)$ for all $(Y|X)$ in $O(T)$.

In other words, a variable pair is said to be JC-derivable from another variable pair if the former may be obtained from the latter through a conditioning followed by a projection — or equivalently, on account of (8), through a projection followed by a conditioning. Note that if $(V|U) \preceq (Y|X)$, then the variables W and Z satisfying the equation in Definition 1 are uniquely determined by

$$W = U \setminus X \text{ and } Z = Y \setminus (V \cup U).$$

Also note this bi-conditional

$$(V|U) \preceq (Y|X) \text{ if and only if } V \cup U \subseteq Y \cup X \text{ and } U \supseteq X \tag{12}$$

which is readily proved and offers a useful characterization of the relation just defined. A still simpler characterization is expressed by the following formula, which makes use of the set-theoretic labels in Figure 2, with $V = B \cup F \cup G$, $U = D \cup E \cup H$, $Y = A \cup E \cup F$, and $X = C \cup G \cup H$:

$$(V|U) \preceq (Y|X) \text{ if and only if } B \cup C \cup D \cup G = \varnothing. \tag{13}$$

The relation introduced with Definition 1 endows its domain with a regular algebraic organization.

Proposition 1. *The relation $\preceq$ is a partial order over the set $\widetilde{O}(T) = O(T) \cup \{\perp\}$. It organizes this set as a lattice, whose supremum and infimum are the pair $(T|\varnothing)$ and the term $\perp$, respectively, and whose join*

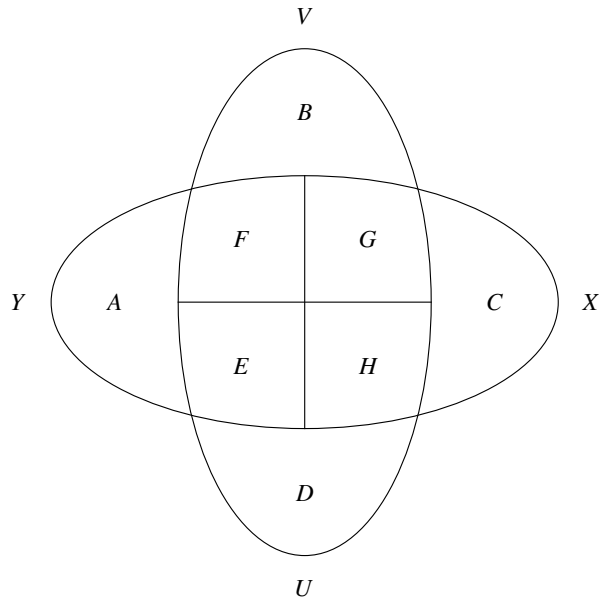

Figure 2. Quatrefoil representing a crossing between two generic variable pairs $(V|U)$ and $(Y|X)$. Some of the eight parts could be empty.

and meet operations are determined by the following equations, for all $(V|U)$ *and* $(Y|X)$ *in* $O(T)$:

$$(V|U) \vee (Y|X) = (V \cup Y \cup (U + X)|U \cap X) \tag{14}$$

with $U + X = (U \setminus X) \cup (X \setminus U)$;

$$(V|U) \wedge (Y|X) = (V \cap Y|U \cup X) \text{ or } = \bot \text{ depending on whether} \tag{15}$$

the conditions $V \cap Y \neq \varnothing$, $U \subseteq Y \cup X$, *and* $X \subseteq V \cup U$ *are or are not jointly true.*

Proof. In the light of characterization (12), it directly appears that reflexivity, transitivity, and antisymmetry of the relation $\preceq$ follow from the homonymous properties of the set-theoretic inclusion $\subseteq$. Thus, $(\widetilde{O}(T), \preceq)$ is a poset (partially ordered set) having $(T|\varnothing)$ as its supremum and $\bot$ as its infimum (concluding statement in Definition 1). Let us consider any two members $(V|U)$ and $(Y|X)$ of the set $O(T)$. We develop our argument concerning their join and meet in three stages. *First* we note that $(V \cup Y \cup (U + X)|U \cap X)$ is itself a member of $O(T)$ (the intersection between left variable and right variable in the pair is empty) and both $(V|U)$ and $(Y|X)$ are JC-derivable from it (according to (12)). Furthermore, if $(Z|W)$ is any member of $O(T)$ such that $(V|U) \preceq (Z|W)$ and $(Y|X) \preceq (Z|W)$, then ($V \cup U \subseteq Z \cup W$ and $U \supseteq W$) and ($Y \cup X \subseteq Z \cup W$ and $X \supseteq W$) (again because of (12)), so that ($V \cup Y \cup (U + X) \subseteq Z \cup W$ and $U \cap X \supseteq W$), which implies $(Y \cup V \cup (U + X)|U \cap X) \preceq (Z|W)$. Thus, $(V \cup Y \cup (U + X)|U \cap X)$ is the least upper bound of $(V|U)$ and $(Y|X)$ in the poset, which proves the equation concerning the join. At a *second stage*, let us suppose that $(V|U)$ and $(Y|X)$ satisfy the three conditions

$$V \cap Y \neq \varnothing, \quad U \subseteq Y \cup X, \quad X \subseteq V \cup U \tag{16}$$

and consider the variable pair $(V \cap Y|U \cup X)$. It is seen that this pair is a member of $O(T)$ (in particular, $V \cap Y \neq \varnothing$ is the first hypothesis in (16)), and is JC-derivable both from $(V|U)$ (in particular, $(V \cap Y) \cup U \cup X \subseteq V \cup U$ is ensured by the third hypothesis in (16)) and from $(Y|X)$ (for similar reasons). Furthermore, if $(Z|W)$ is any member of $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then ($Z \cup W \subseteq V \cup U$ and $W \supseteq U$) and ($Z \cup W \subseteq Y \cup X$ and $W \supseteq X$), so that $Z \cup W \subseteq (V \cup U) \cap (Y \cup X)$ and $W \supseteq U \cup X$. But the second and third hypotheses in (16) imply $(V \cup U) \cap (Y \cup X) = (V \cap Y) \cup (U \cap Y) \cup (V \cap X) \cup (U \cap X) = (V \cap Y) \cup U \cup X$. Therefore, $Z \cup W \subseteq (V \cap Y) \cup U \cup X$ and $W \supseteq U \cup X$, which means $(Z|W) \preceq (V \cap Y|U \cup X)$. Because of the genericity of $(Z|W)$, this proves that $(V \cap Y|U \cup X)$ is the greatest lower bound of $(V|U)$ and $(Y|X)$ in the poset, and confirms the stated formula for the meet. At a *third stage*, we refer to any two members $(V|U)$ and $(Y|X)$ of $O(T)$ that falsify some of the conditions in (16) and prove that there cannot exist any member $(Z|W)$ of $O(T)$ such that both $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, so that $\perp$ is the only common lower bound of $(V|U)$ and $(Y|X)$ in the poset, which means $(V|U) \wedge (Y|X) = \perp$. Suppose the first condition in (16) is false, that is $V \cap Y = \varnothing$ is true. Should a pair $(Z|W)$ exist in $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then (12) combined with $V \cap U = \varnothing = Y \cap X$ would imply $V \cap Y \supseteq Z$, so that $Z = \varnothing$, which contradicts the assumption $(Z|W) \in O(T)$. Next, suppose the second condition in (16) is false, that is $U \setminus (Y \cup X) \neq \varnothing$ is true, so that there is some non-empty variable $S \subseteq U \setminus (Y \cup X)$. Should a pair $(Z|W)$ exist in $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then we would have $S \subseteq W$ (because $W \supseteq U$) and not($S \subseteq W$) (because $W \subseteq Y \cup X$), which of course is contradictory. The same argument may be applied when the third condition in (16) is false. □

Using the labels in Figure 2, the equations that specify joins and meets in the lattice $\widetilde{O}(T)$ can be written as

$$(V|U) \vee (Y|X) = (A \cup B \cup C \cup D \cup E \cup F \cup G|H) \tag{17}$$
$$(V|U) \wedge (Y|X) = (F|E \cup G \cup H) \text{ if } F \neq \varnothing \text{ and } C \cup D = \varnothing.$$

The atoms in the lattice are the pairs $(V|U)$ such that $|V| = 1$, that is, the left-hand component is an elementary variable. Thus, if $n = |T|$, then there are $n2^{n-1}$ atoms. It is readily proved that any member of $O(T)$ is expressible as the join of suitably chosen atoms and that the lattice is rankable, the rank of any pair $(V|U)$ being the cardinality $|V|$ of its left-hand component. Figure 3 illustrates the concept by showing three-dimensional Hasse diagrams of three lattices of variable pairs. In these diagrams, each backward (resp., downward) line represents a projection operation $J[(Y|X), W]$ (resp., a conditioning operation $C[(Y|X), W]$) with $|W| = 1$.

The relation $\preceq$, which organizes the set $\widetilde{O}(T)$ as a lattice, has been defined in terms of projection J and conditioning C as specified by (2) and (3). Some additional comments are in order concerning promotion M as specified by (4). First, if $(V|U)$ is a promoter of $(Y|X)$ (i.e., $V \cup U = X$), then the two variable pairs are $\preceq$-incomparable (because $Y \cap V = \varnothing$) and the result of their promotion equals their join, that is

$$M[(Y|X), (V|U)] = (Y \cup V|U) = (Y|X) \vee (V|U).$$

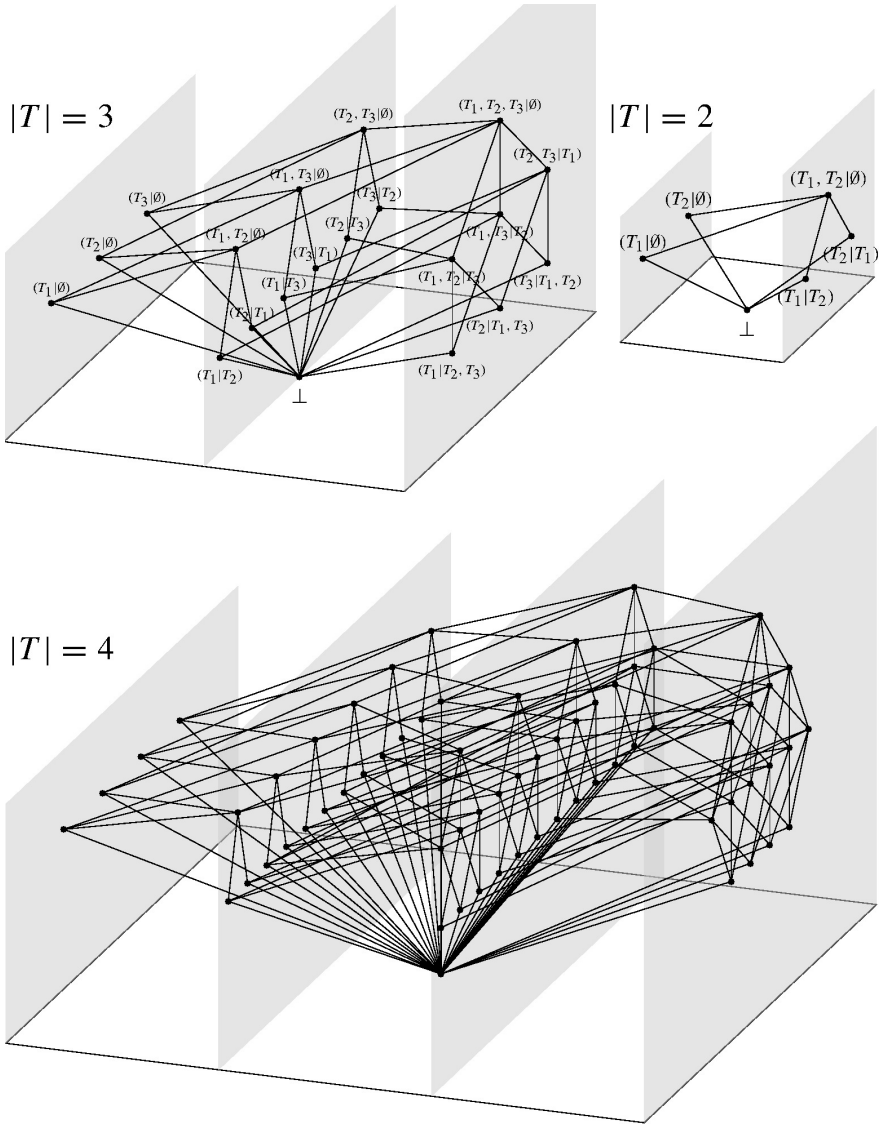

Figure 3. Three-dimensional Hasse diagrams of the lattices $\widetilde{O}(T)$ for $|T| = 2$ (top right), $|T| = 3$ (top left), and (bottom) $|T| = 4$.

Thus, promotion M within a lattice $\widetilde{O}(T)$ is tantamount to a part of the join operation. Second, if $(V|U) \prec (Y|X)$, then one or two promotions are enough for ascending from $(V|U)$ to $(Y|X)$ within the lattice. Specifically, in a writing justified by (9) (associativity of M), the following equation holds true:

$$(Y|X) = (Y \setminus (V \cup U)|V \cup U)\ M\ (V|U)\ M\ (Y \cap U|X). \tag{18}$$

Indeed, if $(V|U) \preceq (Y|X)$ but not $(Y|X) \preceq (V|U)$ (which is the meaning of the hypothesis $(V|U) \prec (Y|X)$), then using the labels in Figure 2 we obtain $B \cup C \cup D \cup G = \varnothing$ but $A \cup E \neq \varnothing$ (because of (13) and

its rewrite characterizing $(Y|X) \preceq (V|U)$), so that $V = F$, $U = E \cup H$, $Y = A \cup E \cup F$, and $X = H$, and (18) may be rewritten as $(A \cup E \cup F|H)$=$(A|E \cup F \cup H)\ M\ (F|E \cup H)\ M\ (E|H)$, which is true according to the definition of the M operation. Note that if either $A = \varnothing$ or $E = \varnothing$, then either the first or the third operand in the right hand side of (18) would be the infimum $\perp$ in the lattice and could be ignored (consistently with (5)), so that one single application of M would be enough for ascending from $(V|U)$ to $(Y|X)$. Third, as a special case of (18) we note the following equation:

$$(V|U) \vee (Y|X) = ((Y \cup X) \setminus (V \cup U)|V \cup U)\ M\ (V|U)\ M\ (U \setminus X|U \cap X).$$

It can be directly proved by noting that the indicated double promotion gives the result $((Y \cup X) \setminus (V \cup U) \cup V \cup (U \setminus X)|U \cap X)$, that is $(A \cup B \cup C \cup D \cup E \cup F \cup G|H)$ by the labeling in Figure 2, and this variable pair is precisely the join $(V|U) \vee (Y|X)$ according to (17). The equation thus proved supplements the first comment in this paragraph, by showing that single and double promotions are enough to simulate the *whole* of the join operation within any lattice of variable pairs.

3. Lattice of probability kernels

In this section we return to probability kernels mentioned in the Introduction, in order to show how the results obtained in discussing variable pairs in the preceding section may conveniently be applied to them.

In the first step, we present the following formulas, which define projection J, conditioning C, and promotion M as operations on probability kernels:

$$J[p(W \cup Z|X), W] = p(Z|X) \tag{19}$$
$$\text{in which } p(z|x) = \sum_{w \in W^\circ} p(w, z|x) \text{ for all } (x, z) \in (X, Z)^\circ$$

$$C[p(W \cup Z|X), Z] = p(W|Z \cup X) \tag{20}$$
$$\text{in which } p(w|z, x) = \frac{p(w, z|x)}{\sum_{w' \in W^\circ} p(w', z|x)} \text{ for all } (x, w, z) \in (X, W, Z)^\circ$$

$$M[p(Y|V \cup U), p(V|U)] = p(Y \cup V|U) \tag{21}$$
$$\text{in which } p(y, v|u) = p(y|v, u) \cdot p(v|u) \text{ for all } (u, v, y) \in (U, V, Y)^\circ.$$

These formulas generalize the rules mentioned in the Introduction and correspond to operations ordinarily performed on conditional probabilities in Bayesian computations (Bernardo & Smith, 2000, pp. 127–130; Koski & Noble, 2009, pp. 53–57). Here we assume that X, W, and Z — as well as U, V, and Y — are (possibly multiple) variables that are disjoint from one another, and that $p(W \cup Z|X)$, $p(Y|V \cup U)$, and $p(V|U)$ are probability kernels acting on them. Formulas (19) and (20), in the given writing, apply when the kernel $p(W \cup Z|X)$ is of discrete type; similar formulas, with Σ replaced by $\int$, are suitable for kernels of continuous type. Of course, the specification of the term $p(w|z, x)$ in formula (20) is acceptable only for any point (x, w, z) such that the denominator in the fraction is non-null, that is, the value $p(z|x)$ resulting from projection is positive. Furthermore, the kernels $p(Y|V \cup U)$ and $p(V|U)$ in formula (21) are here assumed to be of the same measure-theoretic type, that is, either both of them

are families of mass functions (on domains Y° and V°, respectively), or both are families of density functions[2]. Lastly, it is readily seen that there is correspondence between the stated operations on kernels and the operations on variable pairs defined by (2)–(4). For example, if $p(Z|X) = J[p(Y|X), W]$ then $(Z|X) = (Y \setminus W|X) = J[(Y|X), W]$, and similarly for the conditioning and promotion operations. Also it is easily proved that the (6)–(11) still hold true when they are rewritten in terms of probability kernels and operations on these.

The second step in this section is about the relation $\preceq$ specified by Definition 1, which may be recast in terms of probability kernels as follows:

$$\begin{aligned} &p(V|U) \text{ is } JC\text{-derivable from } p(Y|X) \text{ (notation } p(V|U) \preceq p(Y|X)) \\ &\text{if } p(V|U) = J[C[p(Y|X), W], Z] \text{ for some } W, Z \subseteq Y \text{ such that } W \cap Z = \varnothing. \end{aligned} \tag{22}$$

Just as with the relation $\preceq$ between variable pairs, this relation between kernels is a partial order. Indeed, it is reflexive, as W and Z in (22) could be the empty variable. It is transitive, by virtue of properties (6)–(8) as referred to the J and C operations on kernels. It is antisymmetric, because if $p(V|U) \preceq p(Y|X)$ and $p(Y|X) \preceq p(V|U)$, then also $(V|U) \preceq (Y|X)$ and $(Y|X) \preceq (V|U)$, so that $(V|U) = (Y|X)$, which combined with $p(V|U) \preceq p(Y|X)$ implies $p(V|U) = p(Y|X)$. Note that, in comparing any two kernels $p(V|U)$ and $p(V|W \cup U)$, it could turn out that

$$p(v|u) = p(v|w, u) \text{ for all } (u, v, w) \in (U, V, W)^\circ \tag{23}$$

which would mean that V and W are "conditionally independent" given U (Dawid, 1979, p. 3). In that event, $p(V|W \cup U)$ would amount to a replica of $p(V|U)$, and we would accept $p(V|U) \preceq p(V|W \cup U)$ as a true sentence, for a reason similar to accepting $p(V|U) \preceq p(V|U)$ as a true sentence.

In order to relate Proposition 1 to probability kernels, we need to refer to a complete collection of mutually consistent kernels. Let a full kernel $p(T) = p(T|\varnothing)$ be given, that is, one single probability function over the range T° of the assumed full variable T. Then, in correspondence to any variable pair $(Y|X) \in O(T)$, we may JC-derive a kernel $p(Y|X)$ from $p(T|\varnothing)$ by applying the conditioning operation (20) relative to the variable X (i.e., variable X is transferred from the left to the right side of the bar) and then applying the projection operation (19) relative to the variable $T \setminus (Y \cup X)$ (i.e., variable $T \setminus (Y \cup X)$ is canceled from $T \setminus X$, so that precisely Y is what remains on the left side of the bar). By doing so for each of the variable pairs in $O(T)$, a complete collection of kernels is obtained, here denoted by $P(T)$ and thus formally defined:

$$P(T) = \big\{p(Y|X) : p(Y|X) = J[C[p(T|\varnothing), X], T \setminus (Y \cup X)] \text{ for } (Y|X) \in O(T)\big\}.$$

[2]This is a limitation of the M operation as understood in this paper, which may be overcome by setting the concept of probability kernel in measure-theoretic terms. Kernels $p(Y|V \cup U) = (p(Y|v, u) : v \in V^\circ, u \in U^\circ)$ and $p(V|U) = (p(V|u) : u \in U^\circ)$ may generally be families of Radon-Nikodym derivatives (of probability distributions) with respect to reference measures (possibly of different kinds) μ and ν on the spaces Y° and V°, respectively (Billingsley, 1995, pp. 439–440; Pollard, 2002, pp. 84, 119). Hence, for each $u \in U^\circ$ the product function $p(Y \cup V|u) = (p(y|v, u) \cdot p(v|u) : y \in Y^\circ, v \in V^\circ)$ in turn is a Radon-Nikodym derivative (of a probability distribution) with respect to the product measure $\mu \times \nu$ on the product space $Y^\circ \times V^\circ$, and the promoted kernel $p(Y \cup V|U) = (p(Y \cup V|u) : u \in U^\circ)$ is the whole collection of these derivatives.

The kernels in the collection $P(T)$ are mutually consistent as they originate from the same "parent" $p(T)$ — indeed, $P(T)$ is the collection of all kernels that are $\preceq$-dominated by the assumed full distribution $p(T)$. In addition, let $\widetilde{P}(T)$ stand for $P(T) \cup \{\sharp\}$, where the term $\sharp$ — here called the *null kernel* — is assumed to be lower in the order $\preceq$ than all members of $P(T)$ and represents "fictitious kernels" $p(\varnothing|X)$ for $X \subseteq T$[3]. The one-to-one correspondence between variable pairs in $O(T)$ and kernels in $P(T)$ — and between the terms $\perp$ and $\sharp$ — is an isomorphism between the ordered sets $(\widetilde{O}(T), \preceq)$ and $(\widetilde{P}(T), \preceq)$. Proposition 1 shows that the former set is a lattice, so that also the latter is a lattice. The assumed full kernel $p(T)$ and the null kernel $\sharp$ are the supremum and the infimum in the lattice $\widetilde{P}(T)$. The kernels $p(Y|X)$ in which $|Y| = 1$ are the atoms. The join and meet operations are defined by formulas that duplicate (14) and (15) in terms of probability kernels.

The preceding argument has been cast in a top-down perspective, as a full kernel $p(T)$ has been assumed available, from which a complete collection $P(T)$ of mutually consistent kernels may be derived. But, in the construction of probabilistic models, a perspective somewhat opposite to this is actually taken. Indeed, in constructing a model, kernels of low rank are first specified — that is, kernels $p(Y|X)$ in which Y is a variable of small cardinality, possibly an elementary variable, for simplicity. These are the building blocks of the model, from which other kernels of higher rank are obtained — by combining the building blocks through promotion or multiplication under suitable assumptions of stochastic independence — and then other lower rank kernels can be JC-derived for the necessities of the modeling (Koller & Friedman, 2009, pp. 4–5).

Illustrations of this circumstance are offered by hierarchical Bayesian models in statistics. Let us consider, for example, the following assignment formulas, in which D is an observable random variable (it could be the mean of a sample of data), whose distribution is assumed to involve a location parameter μ and a precision parameter λ, which themselves are conceived of as random variables with distributions depending on hyper-parameters ν, ξ, α, and β, these being provided with definite hyper-prior distributions (a model compatible with Bernardo & Smith, 2000, p. 440):

$$\begin{aligned} &D \sim \text{Normal}(\mu, \lambda), \quad \mu \sim \text{Normal}(\nu, \xi), \quad \lambda \sim \text{Gamma}(\alpha, \beta), \\ &\nu \sim \text{Uniform}(0, 50), \quad \xi \sim \text{Uniform}(0, 1), \quad \alpha \sim \text{Uniform}(0, 1), \quad \beta \sim \text{Uniform}(0, 1). \end{aligned} \tag{24}$$

In the terms we are using in this paper, these formulas specify seven primitive kernels $p(D|\mu, \lambda)$, $p(\mu|\nu, \xi)$, $p(\lambda|\alpha, \beta)$, $p(\nu|\varnothing)$, $p(\xi|\varnothing)$, $p(\alpha|\varnothing)$, and $p(\beta|\varnothing)$, which express postulates in the model (their variable pairs are atoms in a lattice, as the conditioned variables are one-dimensional). For purposes of statistical inference, the modeler may be interested in determining the kernel $p(\mu|D)$ that is implied by the assumed postulates, that is, the family of posterior distributions of the parameter μ given the observable quantity D. But for determining such low rank kernel (itself an atom) one must first ascend from the given primitive kernels to the top kernel $p(D, \mu, \lambda, \nu, \xi, \alpha, \beta|\varnothing)$ and then descend from this to $p(\mu|D)$ through suitable conditioning and projections. Of course, such a procedure is only plausible if a kernel $p(D, \mu, \lambda, \nu, \xi, \alpha, \beta|\varnothing)$ that $\preceq$-dominates all seven primitive kernels really exists (and is

[3] In numerical operations, the null kernel $\sharp$ acts as the number 1, which is the neutral term of multiplication (cf. Studený, 2005, p. 20).

reachable from these by ascending operations, possibly supported by suitable independence assumptions). In general, however, if the primitive kernels in a model are specified separately from one another, there is no a priori guarantee that there is a "consensus kernel" covering all of them, so that the stated problem might fail to have a solution.

In addressing this topic, use must be made of the concept of "compatibility" as understood in the discussions concerning conditional specified distributions and, more generally, conditionally specified statistical models (Arnold, Castillo, & Sarabia, 1999, 2001). The concept can consistently be framed within the theory developed so far in this paper.

Definition 2. Given kernels $p(Y_1|X_1), \ldots, p(Y_m|X_m)$ are *compatible* with one another if there is a kernel $p(Z|W)$ from which all of them are JC-derivable, that is, $p(Y_i|X_i) \preceq p(Z|W)$ for all $i = 1, \ldots, m$.

Note that if $p(Z|W)$ is a kernel that satisfies this condition, then $(Y_i|X_i) \preceq (Z|W)$ for all $i = 1, \ldots, m$, so that $(Y_1|X_1) \vee \cdots \vee (Y_m|X_m) \preceq (Z|W)$. Thus, for verifying whether m given kernels are mutually compatible, the standard approach would be to verify whether on the *join* of their variable pairs a kernel may be constructed from which all of them can be JC-derived. Also note that this compatibility relation fails to be universally transitive. For example, if X and Y are disjoint discrete variables, then any kernel $p(X|\varnothing)$ is compatible both with any kernel $p(Y|X)$ and with any kernel $p(Y|\varnothing)$, but these two could be incompatible with each other — certainly they are compatible if $p(Y|\varnothing) = J[M[p(Y|X), p(X|\varnothing)], X]$.

For applications, one wants to have conditions that are easily testable and sufficient to ensure compatibility. The next proposition provides such a condition, which is rather general, as it *only* concerns the *variable pairs* on which the kernels are acting. No mention is made of the specific form of the probability functions in the kernels.

Proposition 2. *Let* $(p(Y_1|X_1), \ldots, p(Y_m|X_m))$ *be a list of* $m \geq 2$ *probability kernels of the same type* (*i.e.*, *either all of discrete type*, *or all of continuous type*) *such that their variable pairs satisfy this condition*:

$$Y_i \cap (Y_{i-1} \cup X_{i-1} \cup \cdots \cup Y_1 \cup X_1) = \varnothing, \text{ for all } i = 2, \ldots, m. \tag{25}$$

Then the m kernels are compatible.

Proof. The proof is by induction on the number $m \geq 2$ of kernels. *First step*: For $m = 2$, let any two variable pairs $(Y_1|X_1) = (Y|X)$ and $(Y_2|X_2) = (V|U)$ be given such that $V \cap (Y \cup X) = \varnothing$, that is $F \cup G = \varnothing$ in the terms of Figure 2, so that $(Y|X) \vee (V|U) = (A \cup B \cup C \cup D \cup E|H)$ according to (17). Let $p(Y|X) = p(A \cup E|C \cup H)$ and $p(V|U) = p(B|D \cup E \cup H)$ be arbitrary kernels of the same measure-theoretic type on the indicated variable pairs. We may extend these kernels into $p(A \cup E \cup D|C \cup H)$ and $p(B|A \cup C \cup D \cup E \cup H)$ by setting

$$p(a, e, d|c, h) = p(a, e|c, h) \cdot p(d), \text{ for all } (a, e, d, c, h) \in (A, E, D, C, H)^\circ$$
$$p(b|a, c, d, e, h) = p(b|d, e, h), \text{ for all } (b, a, c, d, e, h) \in (B, A, C, D, E, H)^\circ$$

where $p(D)$ is a freely chosen kernel of the same type as the two given kernels. Then we may combine the extended kernels by promotion to obtain the following kernel on the join $(Y|X) \vee (V|U)$:

$$p(A \cup B \cup C \cup D \cup E|H) = p(B|A \cup C \cup D \cup E \cup H) \; M \; p(A \cup E \cup D|C \cup H) \; M \; p(C|H)$$

where $p(C|H)$ is itself a freely chosen kernel. Note that the double promotion in this equation is syntactically regular, as the third kernel is a promoter of the second, and in turn this is a promoter of the first. Kernel $p(A \cup E|C \cup H)$ is derivable from $p(A \cup E \cup D|C \cup H)$ by projection, and this is derivable from $p(A \cup B \cup C \cup D \cup E|H)$ because of (10) and (11). Furthermore, the remark associated to (23) ensures that $p(B|D \cup E \cup H)$ is derivable from $p(B|A \cup C \cup D \cup E \cup H)$, and this is derivable from $p(A \cup B \cup C \cup D \cup E|H)$ because of (10). Therefore, $p(Y|X) = p(A \cup E|C \cup H)$ and $p(V|U) = p(B|D \cup E \cup H)$ are compatible kernels, as there exists a kernel from which both of them are derivable. Note that $p(A \cup B \cup C \cup D \cup E|H)$ does not involve any variable besides those involved in $p(A \cup E|C \cup H)$ or in $p(B|D \cup E \cup H)$. *Inductive step*: Let us now consider any list $(p(Y_1|X_1), \ldots, p(Y_{m-1}|X_{m-1}), p(Y_m|X_m))$ of $m > 2$ kernels whose variable pairs comply with condition (25), and suppose (as inductive hypothesis) that the first $m - 1$ members in the list are compatible with one another, so that there is a kernel $p(Z|W)$ from which all of them are derivable. Based on the remark that concludes the first step of the current proof, we may presume $Z \cup W \subseteq Y_{m-1} \cup X_{m-1} \cup \cdots \cup Y_1 \cup X_1$, so that $Y_m \cap (Z \cup W) = \varnothing$ because of hypothesis (25). Thus, the conditions are satisfied that make it possible to apply the argument in the first step of the current proof to the kernels $p(Z|W)$ and $p(Y_m|X_m)$. The argument ensures the existence of a kernel from which both $p(Z|W)$ (hence $p(Y_1|X_1), \ldots, p(Y_{m-1}|X_{m-1})$) and $p(Y_m|X_m)$ are derivable. Therefore, the m kernels are all compatible with one another. □

The proposition thus proved directly shows the internal consistency of the Bayesian model specified in (24). Indeed, if the primitive kernels in the model are listed in the order

$$(p(\beta|\varnothing),\ p(\alpha|\varnothing),\ p(\xi|\varnothing),\ p(\nu|\varnothing),\ p(\lambda|\alpha, \beta),\ p(\mu|\nu, \xi),\ p(D|\mu, \lambda)),$$

then it appears that the conditioned variable in each kernel falls out of the collection of the variables involved in the kernels preceding that kernel in the list, so that condition (25) for mutual compatibility of the kernels is satisfied. More generally, suppose that $(T_{r(1)}, \ldots, T_{r(n)})$ is an ordering of the elementary variables that form the full variable $T = \{T_1, \ldots, T_n\}$ of a probabilistic model, and that $(T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n)$ are variable pairs — specifically, atoms in the lattice $\widetilde{O}(T)$ — such that $X_i \subseteq T_{r(1)} \cup \ldots \cup T_{r(i-1)}$, for each $i = 1, \ldots, n$ (in particular, $X_1 = \varnothing$). Then Proposition 2 implies, as a corollary, that *any* kernels $p(T_{r(1)}|X_1), \ldots, p(T_{r(n)}|X_n)$ of the same measure-theoretic type (e.g., all discrete or all continuous kernels) on those variable pairs are compatible with one another, and such kernels uniquely determine (through multiple promotion, possibly supported by assumptions of stochastic independence) a distribution $p(T)$ on the full variable of the model. Thus, possible kernels on the atom variable pairs $(T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n)$ form a system of independent and minimum-rank generators of possible full distributions for the model.

The corollary now highlighted corresponds to a key result in the theory of Bayesian networks. Indeed, if the elementary variables in a collection $T = \{T_1, \ldots, T_n\}$ are represented as nodes of an acyclic directed graph (DAG), then a list $((T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n))$ of variable pairs with the stated characteristics can be formed, in which $(T_{r(1)}, \ldots, T_{r(n)})$ is a suitable permutation of T and X_i (for $i = 1, \ldots, n$) is the set of "parents" of the variable $T_{r(i)}$ in the graph. Therefore, if $p(T_{r(1)}|X_1), \ldots, p(T_{r(n)}|X_n)$ are the elementary probability kernels postulated in the Bayesian network, then by virtue of the stated corollary a joint "consensus" distribution $p(T)$ does exist and this may be uniquely inferred (also thanks to the

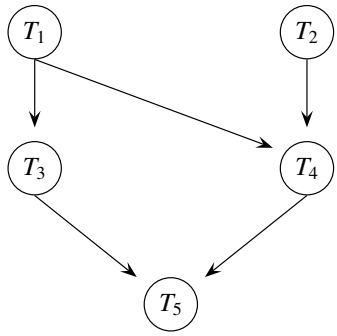

Figure 4. The DAG of a Bayesian network on five elementary variables (from Kjærulff & Madsen, 2008, p. 13).

conditional independences represented in the DAG) through the "chain rule", which corresponds to multiple promotion in our terms (Pearl, 1988, pp. 119–120; Darwiche, 2009, pp. 57–58). The DAG in Figure 4 offers the basis for an illustration of the stated property. By associating to each elementary variable the set of its parents in the graph — that is, the variables emitting an arrow towards that variable — the following list of variable pairs is obtained:

$$(T_1|\varnothing),\ (T_2|\varnothing),\ (T_3|T_1),\ (T_4|T_1,T_2),\ (T_5|T_3,T_4).$$

It is seen that, in this ordering, the variable pairs do comply with condition (25) (a circumstance ultimately due to the acyclic character of the graph), so that Proposition 2 ensures mutual compatibility among the probability kernels that a modeler may specify in defining a Bayesian network on the DAG:

$$p(T_1|\varnothing),\ p(T_2|\varnothing),\ p(T_3|T_1),\ p(T_4|T_1,T_2),\ p(T_5|T_3,T_4). \tag{26}$$

Furthermore, the DAG represents a system of conditional stochastic independences, which are assumed true by the modeler. For example, it implies that the variable T_5 is assumed to be conditionally independent of the variable $\{T_1, T_2\}$ given the variable $\{T_3, T_4\}$, this being the set of parents of T_5. Because of such independences encoded in the hypothesized graph, the specification (by the modeler) of the kernels (26) directly entails (on account of (23)) the specification of these kernels:

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

It is seen that, in this ordering (which for convenience is the opposite of that in (26)), each kernel is a regular promoter of its immediate predecessor in the list. Thus, multiple promotion can be applied, whose result will be the full distribution $p(T_1, T_2, T_3, T_4, T_5)$ the existence of which is ensured by the configuration of the variable pairs owing to Proposition 2. Such multiple promotion amounts to an ascension from five atoms in a lattice of kernels to their join (which, in this special case, equals the supremum of the lattice itself).

4. Order of variable pairs and compatibility of kernels

In this section, we present a result that illustrates the diagnostic power of the order $\preceq$ between variable pairs as concerns a special aspect of compatibility between probability kernels. More precisely, we will show that, for all variable pairs $(V|U)$ and $(Y|X)$, one has $(V|U) \preceq (Y|X)$ if and only if *each* kernel $p(Y|X)$ on the latter pair *uniquely determines* a kernel $p(V|U)$ on the former. In other words, if $p(Y|X)$ is given and $(V|U) \preceq (Y|X)$, then there exists one single kernel on $(V|U)$ compatible with it, whereas if not $(V|U) \preceq (Y|X)$, then there exist different kernels on $(V|U)$ compatible with the same $p(Y|X)$. As in the preceding section, implicit in the following arguments is the assumption that the kernels to be combined or compared are of the same measure-theoretic type (either discrete or continuous).

In proving the indicated result, use will be made of one further operation on kernels, called *multiplication*, which is defined as follows (cf. Koski & Noble, 2009, pp. 55–56):

$$\begin{aligned}
&\text{for all } p(Y|X) \text{ and } p(Z|W) \text{ such that } Y \cap (Z \cup W) = \varnothing = Z \cap (Y \cup X) \\
&p(Y|X) \times p(Z|W) = p(Y \cup Z|X \cup W) \\
&\text{in which } p(y, z|t, u, v) = p(y|t, u) \cdot p(z|u, v) \\
&\text{for all } (y, z) \in (Y, Z)^\circ \text{ and } (t, u, v) \in (X \setminus W, X \cap W, W \setminus X)^\circ.
\end{aligned} \tag{27}$$

This formula is applicable also to cases in which $W = \varnothing$ (so that $p(Z|W) = p(Z|\varnothing)$ is one single probability function on the range Z°) or $Z = \varnothing$ (so that $p(Z|W) = p(\varnothing|W)$ stands for the null kernel $\sharp$; see footnote 3). In these special cases, the definition becomes

$$\begin{aligned}
&p(Y|X) \times p(Z|\varnothing) = p(Y \cup Z|X) \\
&\qquad \text{in which } p(y, z|x) = p(y|x) \cdot p(z) \text{ for all } (y, z, x) \in (Y, Z, X)^\circ \\
&p(Y|X) \times p(\varnothing|W) = p(Y|X \cup W) \\
&\qquad \text{in which } p(y|x, w) = p(y|x) \text{ for all } (y, x, w) \in (Y, X, W)^\circ.
\end{aligned}$$

It is easily shown that the product $p(Y|X) \times p(Z|W)$ defined by (27) is itself a probability kernel; more specifically, it is a family (indexed by $(X \cup W)^\circ$) of probability functions on $(Y, Z)^\circ$. Equally it can be shown that multiplication is an associative and commutative operation, and for every variable $S \subseteq Y$ it satisfies these equations:

$$J[p(Y|X) \times p(Z|W), S] = (J[p(Y|X), S]) \times p(Z|W) \tag{28}$$

$$C[p(Y|X) \times p(Z|W), S] = (C[p(Y|X), S]) \times p(Z|W). \tag{29}$$

In other words, a kind of commutativity holds between multiplication $\times$ on the one hand and projection J and conditioning C on the other hand.

Now we state and prove the main result in this section.

Proposition 3. *Let $(V|U)$ and $(Y|X)$ be any two non-null members of a lattice $\widetilde{O}(T)$ of variable pairs. Then $(V|U) \preceq (Y|X)$ if and only if for all kernels $p(V|U)$, $p(Y|X)$, $q(V|U)$, and $q(Y|X)$ such that the*

first two are compatible, *and also the other two are compatible*, *the equality* $p(Y|X) = q(Y|X)$ *implies the equality* $p(V|U) = q(V|U)$.

Proof. The "only if" part of this proposition is easily shown, on considering that if $(V|U) \preceq (Y|X)$, then the stated compatibility hypotheses imply $p(V|U) \preceq p(Y|X)$ and $q(V|U) \preceq q(Y|X)$, that is

$$p(V|U) = J[C[p(Y|X), U \setminus X], Y \setminus (V \cup U)] \text{ and } q(V|U) = J[C[q(Y|X), U \setminus X], Y \setminus (V \cup U)]$$

so that the equality $p(Y|X) = q(Y|X)$ obviously implies the equality $p(V|U) = q(V|U)$. To prove the "if" part is tantamount to proving that

$$\begin{aligned} &\text{if not } (V|U) \preceq (Y|X) \\ &\text{then there are kernels } p(V|U),\ p(Y|X), q(V|U), \text{ and } q(Y|X) \text{ such that} \\ &p(V|U) \text{ and } p(Y|X) \text{ are compatible, } q(V|U) \text{ and } q(Y|X) \text{ are compatible, and} \\ &p(Y|X) = q(Y|X) \text{ but } p(V|U) \neq q(V|U). \end{aligned} \tag{30}$$

In the terms of Figure 2 and on account of (13), the antecedent "not $(V|U) \preceq (Y|X)$" in this implication means that the condition $B \cup C \cup D \cup G \neq \varnothing$ holds true. Hereafter we separately discuss (using the labels in Figure 2) three cases that exhaustively cover this condition. *Case* $B \cup G \neq \varnothing$. Choose any two probability functions $p(B \cup G)$ and $q(B \cup G)$ that are *different* from each other — such functions do exist, simply because $B \cup G \neq \varnothing$. Then consider any probability function $r(A \cup C \cup D \cup E \cup F \cup H)$, combine it with $p(B \cup G)$ and $q(B \cup G)$ by multiplication, thus obtaining

$$\begin{aligned} p(W) &= p(A \cup B \cup C \cup D \cup E \cup F \cup G \cup H) = p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H) \\ q(W) &= q(A \cup B \cup C \cup D \cup E \cup F \cup G \cup H) = q(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H) \end{aligned}$$

JC-derive the following kernels from $p(W)$

$$\begin{aligned} p(Y|X) &= p(A \cup E \cup F|C \cup G \cup H) = J[C[p(W), C \cup G \cup H], B \cup D] \\ p(V|U) &= p(B \cup F \cup G|D \cup E \cup H) = J[C[p(W), D \cup E \cup H], A \cup C] \end{aligned}$$

and similarly the kernels $q(Y|X)$ and $q(V|U)$ from $q(W)$. Kernels $p(Y|X)$ and $p(V|U)$ are compatible, because they are JC-derived from the *same* $p(W)$, and for a similar reason also $q(Y|X)$ and $q(V|U)$ are compatible. Furthermore, on account of (28) and (29),

$$\begin{aligned} &p(Y|X) = \\ &J[C[p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H), C \cup G \cup H], B \cup D] = \\ &J[C[p(B \cup G), G] \times C[r(A \cup C \cup D \cup E \cup F \cup H), C \cup H], B \cup D] = \\ &J[p(B|G) \times r(A \cup D \cup E \cup F|C \cup H), B \cup D] = \\ &J[p(B|G), B] \times J[r(A \cup D \cup E \cup F|C \cup H), D] = \\ &p(\varnothing|G) \times r(A \cup E \cup F|C \cup H) \end{aligned}$$

and

$$
\begin{aligned}
&p(V|U) = \\
&J[C[p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H),\, D \cup E \cup H],\, A \cup C] = \\
&J[p(B \cup G) \times C[r(A \cup C \cup D \cup E \cup F \cup H),\, D \cup E \cup H],\, A \cup C] = \\
&J[p(B \cup G) \times r(A \cup C \cup F|D \cup E \cup H),\, A \cup C] = \\
&p(B \cup G) \times J[r(A \cup C \cup F|D \cup E \cup H),\, A \cup C] = \\
&p(B \cup G) \times r(F|D \cup E \cup H).
\end{aligned}
$$

Similar procedures show that $q(Y|X) = q(\varnothing|G) \times r(A \cup E \cup F|C \cup H)$ and $q(V|U) = q(B \cup G) \times r(F|D \cup E \cup H)$. Therefore, $p(Y|X) = q(Y|X)$ (because $p(\varnothing|G) = \sharp = q(\varnothing|G)$), but $p(V|U) \neq q(V|U)$ (because $p(B \cup G) \neq q(B \cup G)$ by the initial choice), so that the four kernels do comply with the requirements in the consequent of implication (30). *Case* $B \cup G = \varnothing$ and $D \neq \varnothing$. As $(V|U)$ is presumed different from $\perp$ (the null variable pair), the hypothesis $B \cup G = \varnothing$ implies that $F \neq \varnothing$. It is well known that, for any two (non-empty) disjoint variables F and D, probability functions $p(F \cup D)$ and $q(F \cup D)$ can be constructed such that $p(F) = J[p(F \cup D),\, D]$ and $q(F) = J[q(F \cup D),\, D]$ are *equal*, but $p(F|D) = C[p(F \cup D),\, D]$ and $q(F|D) = C[q(F \cup D),\, D]$ are *different*. Given such functions, by multiplication let us construct the following:

$$
\begin{aligned}
p(W) &= p(A \cup C \cup D \cup E \cup F \cup H) = p(F \cup D) \times r(A \cup C \cup E \cup H) \\
q(W) &= q(A \cup C \cup D \cup E \cup F \cup H) = q(F \cup D) \times r(A \cup C \cup E \cup H)
\end{aligned}
$$

where $r(A \cup C \cup E \cup H)$ is a freely chosen probability function. By applying the same method used above, the following results are obtained:

$$
\begin{aligned}
p(Y|X) &= p(A \cup E \cup F|C \cup H) = J[C[p(W),\, C \cup H],\, D] = p(F) \times r(A \cup E|C \cup H) \\
p(V|U) &= p(F|D \cup E \cup H) = J[C[p(W),\, D \cup E \cup H],\, A \cup C] = p(F|D) \times r(\varnothing|E \cup H)
\end{aligned}
$$

and similarly $q(Y|X) = q(F) \times r(A \cup E|C \cup H)$ and $q(V|U) = q(F|D) \times r(\varnothing|E \cup H)$. Thus, the equality $p(F) = q(F)$ implies $p(Y|X) = q(Y|X)$, and the inequality $p(F|D) \neq q(F|D)$ implies $p(V|U) \neq q(V|U)$, so that the four kernels satisfy what the implication (30) requires. *Case* $B \cup D \cup G = \varnothing$ and $C \neq \varnothing$. A well-known fact, symmetric to that mentioned above, is that for any two (non-empty) disjoint variables F and C, probability functions $p(F \cup C)$ and $q(F \cup C)$ can be constructed such that $p(F) = J[p(F \cup C),\, C]$ and $q(F) = J[q(F \cup C),\, C]$ are *different*, whereas $p(F|C) = C[p(F \cup C),\, C]$ and $q(F|C) = C[q(F \cup C),\, C]$ are *equal*. Drawing on this property, the third case in question can be solved by the same method used for the previous ones, on taking account that $(Y|X) = (A \cup E \cup F|C \cup H)$ and $(V|U) = (F|E \cup H)$ in the presumed situation. □

The proposition thus proved shows that the order relation $\preceq$ between variable pairs constitutes a general criterion for deducibility between probability kernels. For illustrating the meaning of this statement, let us assume $T = \{T_1, T_2, T_3\}$ as the full variable of a model and consider the following three variable pairs

in the lattice $\widetilde{O}(T)$ (see Figure 3 on page 86, top left):

$$(T_1, T_2|T_3),\ (T_1|T_2, T_3),\ (T_3|T_1, T_2).$$

In principle, on each of these variable pairs various (infinitely many) alternative probability kernels may be defined, but suppose that compatibility between kernels is required (cf. Definition 2). We may then ask: given the compatibility requirement, does a deterministic constraint exist between how a kernel on $(T_1, T_2|T_3)$ is chosen and how kernels on $(T_1|T_2, T_3)$ and $(T_3|T_1, T_2)$ may be chosen? Proposition 3 allows us to give the following answers. For *any* possible kernel $p(T_1, T_2|T_3)$ there is *one single* kernel $p(T_1|T_2, T_3)$ compatible with it (and derivable from it by a C operation), because the relation $(T_1|T_2, T_3) \preceq (T_1, T_2|T_3)$ between their variable pairs is true. On the contrary, for *any* possible kernel $p(T_1, T_2|T_3)$ there are *several* kernels $p(T_3|T_1, T_2)$ compatible with it, because the relation $(T_3|T_1, T_2) \preceq (T_1, T_2|T_3)$ between their variable pairs is false. It is remarkable that both answers can be given only considering the candidate variable pairs, irrespectively of the specific kernel chosen for the pair $(T_1, T_2|T_3)$. In general, for any fixed variable pair $(Y|X)$ in a lattice $\widetilde{O}(T)$, we may ask: which are the variable pairs $(V|U)$ such that, for *any* kernel $p(Y|X)$ there exists *exactly one* kernel $p(V|U)$ *compatible* with it? Proposition 3 answers that such variable pairs $(V|U)$ are precisely those such that $(V|U) \preceq (Y|X)$, that is, the members of the ideal generated by the member $(Y|X)$ within the lattice $\widetilde{O}(T)$.

5. Concluding remarks

The motivations for this study arose from the examination of probabilistic models that involve *several* variables and assign a prominent role to *conditional* probability distributions on them. Examples are provided by Bayesian networks, probabilistic graphical models, hierarchical Bayesian models in statistics, and Bayesian and Markov models in experimental sciences, which we mentioned in the Introduction along with few selected references to the corresponding literature. In models of these kinds, distinct *levels* of conditional probabilities are generally involved, in which the role played by any one variable in the system may vary. An elementary illustration of such duplicity or reversibility of roles is provided by the Bayes rule itself, as it intervenes in basic statistical models. Within the equation $p(\theta|D) = p(\theta) \cdot p(D|\theta)/p(D)$ that expresses the rule, the parameter quantity θ is a conditioning variable (placed on the right of the bar) in the likelihood term $p(D|\theta)$, and becomes a conditioned variable (placed on the left of the bar) in the posterior term $p(\theta|D)$. The opposite is true of the variable D representing the data.

With this study, we contribute ideas for a general framework in which the key components of such probabilistic models may be represented and interrelated for analysis. For representing the key components of a model, the comprehensive concept of "probability kernel" has been adopted, characterized as a family of probability functions on one (possible multiple) variable that are indexed by some other (possible multiple) variable. The analysis has been focused on such pairs of multiple variables and on the set-theoretic relations and operations applicable to them. This analysis is bent toward generality, as it is independent of the peculiar characteristics of the probability functions collected in a kernel. Working in this perspective, significant implications of algebraic character have been found, especially relating to the concept of a lattice.

Although focused on the variables, our analysis has meaningful consequences concerning the probability kernels acting on the given variables. Proposition 1 reveals a kind of algebraic structure into which the kernels in a complex probabilistic model may be mapped and interrelated through operations, in a way suggested by the Hasse diagrams in Figure 3. Proposition 2 provides a general criterion of compatibility between kernels, that makes it possible to test whether given low rank kernels may be the building blocks of a consistent probabilistic model. Proposition 3 provides a criterion for finding which kernels are uniquely determined by a given kernel and may thus be unambiguously deduced from it. The criterion is quite general and becomes strengthened when assumptions are introduced that specify the kind of probability functions forming the kernels in a model — for example, deducibility between kernels of Gaussian form.

References

[1] B. C. Arnold, E. Castillo, and J. M. Sarabia, *Conditional specification of statistical models*, Springer, 1999.

[2] B. C. Arnold, E. Castillo, and J. M. Sarabia, "Conditionally specified distributions: an introduction", *Statist. Sci.* **16**:3 (2001), 249–274.

[3] J.-M. Bernardo and A. F. M. Smith, *Bayesian theory*, Wiley, Chichester, 2000.

[4] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, New York, 1995.

[5] A. Blake, P. Kohli, and C. Rother (editors), *Markov random fields for vision and image processing*, MIT Press, Cambridge, MA, 2011.

[6] A. Darwiche, *Modeling and reasoning with Bayesian networks*, Cambridge University Press, 2009.

[7] A. P. Dawid, "Conditional independence in statistical theory", *J. Roy. Statist. Soc. Ser. B* **41**:1 (1979), 1–31.

[8] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed., CRC Press, Boca Raton, FL, 2014.

[9] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference", *Annu. Rev. Psychol.* **55** (2004), 271–304.

[10] U. B. Kjærulff and A. L. Madsen, *Bayesian networks and influence diagrams: a guide to construction and analysis*, Springer, 2008.

[11] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*, MIT Press, Cambridge, MA, 2009.

[12] T. Koski and J. M. Noble, *Bayesian networks: an introduction*, Wiley, Chichester, 2009.

[13] S. L. Lauritzen, *Graphical models*, Oxford Statistical Science Series **17**, Oxford University Press, New York, 1996.

[14] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer, 1993.

[15] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, San Mateo, CA, 1988.

[16] D. Pollard, *A user's guide to measure theoretic probability*, Cambridge Series in Statistical and Probabilistic Mathematics **8**, Cambridge University Press, 2002.

[17] J. N. Rouder, R. D. Morey, and M. S. Pratte, "Bayesian hierarchical models of cognition", pp. 504–551 in *New handbook of mathematical psychology*, vol. 1, edited by W. H. Batchelder et al., Cambridge University Press, 2017.

[18] M. Studený, *Probabilistic conditional independence structures*, Springer, 2005.

Received 2018-09-06. Revised 2019-09-13. Accepted 2019-10-09.

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

msp.org/astat

See inside back cover or msp.org/astat for submission instructions.

Algebraic Statistics (ISSN 2693-3004 electronic, 2693-2997 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

AStat peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
http://msp.org/

Algebraic Statistics

2020 11 : 1