The Annals of Statistics
2017, Vol. 45, No. 3, 1152–1184
DOI: 10.1214/16-AOS1478

TOTAL POSITIVITY IN MARKOV STRUCTURES

BY SHAUN FALLAT[*,1], STEFFEN LAURITZEN[†], KAYVAN SADEGHI[‡,2], CAROLINE UHLER[§,3], NANNY WERMUTH[¶] AND PIOTR ZWIERNIK[‖,4]

University of Regina[*], *University of Copenhagen*[†], *University of Cambridge*[‡], *Massachusetts Institute of Technology*[§], *IST Austria*[§], *Chalmers University of Technology*[¶], *Johannes Gutenberg-University*[¶] *and Universitat Pompeu Fabra*[‖]

We discuss properties of distributions that are multivariate totally positive of order two (MTP_2) related to conditional independence. In particular, we show that any independence model generated by an MTP_2 distribution is a compositional semi-graphoid which is upward-stable and singleton-transitive. In addition, we prove that any MTP_2 distribution satisfying an appropriate support condition is faithful to its concentration graph. Finally, we analyze factorization properties of MTP_2 distributions and discuss ways of constructing MTP_2 distributions; in particular, we give conditions on the log-linear parameters of a discrete distribution which ensure MTP_2 and characterize conditional Gaussian distributions which satisfy MTP_2.

1. Introduction. This paper discusses a special form of positive dependence. *Positive dependence* may refer to two random variables that have a positive covariance, but other definitions of positive dependence have been proposed as well; see [24] for an overview. Random variables $X = (X_1, \dots, X_d)$ are said to be *associated* if $\mathrm{cov}\{f(X), g(X)\} \geq 0$ for any two nondecreasing functions f and g for which $\mathbb{E}|f(X)|$, $\mathbb{E}|g(X)|$ and $\mathbb{E}|f(X)g(X)|$ all exist [13]. This notion has important applications in probability theory and statistical physics; see, for example, [27, 28].

However, association may be difficult to verify in a specific context. The celebrated FKG theorem, formulated by Fortuin, Kasteleyn and Ginibre in [14], introduces an alternative notion and establishes that X are associated if their joint density function is *multivariate totally positive of order 2*: A function f over $\mathcal{X} = \prod_{v\in V} \mathcal{X}_v$, where each $\mathcal{X}_v$ is totally ordered, is *multivariate totally positive of order two* (MTP_2) if

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y) \qquad \text{for all } x, y \in \mathcal{X},$$

Received October 2015; revised May 2016.
[1]Supported in part by an NSERC Discovery Research Grant.
[2]Supported by Grant #FA9550-12-1-0392 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).
[3]Supported by the Austrian Science Fund (FWF) Y 903-N35.
[4]Supported by the European Union Seventh Framework Programme PIOF-GA-2011-300975.
MSC2010 subject classifications. Primary 60E15, 62H99; secondary 15B48.
Key words and phrases. Association, concentration graph, conditional Gaussian distribution, faithfulness, graphical models, log-linear interactions, Markov property, positive dependence.

where $x \wedge y$ and $x \vee y$ denote the elementwise minimum and maximum.

These inequalities are often easier to check. Furthermore, most other known definitions of positive dependence are implied by the MTP_2 constraints; see, for example, [7] for a recent overview. Note that the conditions are on the probabilities or the density and not on other types of traditional measures of dependence. But as we shall see, the above inequality constraints combined with conditional independence restrictions specify positive associations along edges in undirected graphs, named and studied as dependence graphs or concentration graphs; see, for instance, [22, 44].

MTP_2 distributions have also played an important role in the study of ferromagnetic Ising models, that is, distributions of binary variables where all interaction potentials are pairwise and nonnegative. It has been noted in [33] that the block Gibbs sampler is monotonic if the target distribution is MTP_2, and hence particularly efficient in this setting. Bartolucci and Besag [3], Section 5, showed that much of this work can in fact be extended to arbitrary binary Markov fields. See also [11] for an optimization viewpoint. The special case of Gaussian distributions was studied by Karlin and Rinott [21] and very recently by Slawski and Hein [39] from a machine learning perspective.

Consequences of MTP_2 distributions with respect to marginal and mutual independences were studied by Lebowitz [23] and Newman [28]. They showed in particular that independence of two components of a random vector with an MTP_2 distribution is equivalent to a block-diagonal structure in the covariance matrix and that mutual independence of several components can be inferred from a block-diagonal covariance matrix (see also Theorem 3.7 and Theorem 5.5 below). This is remarkable because covariances and correlations are the weakest types of measures of dependence (see [48]); although they identify independence in Gaussian distributions, this is often not the case for other types of distribution.

In this paper, we discuss implications of the MTP_2 constraints for conditional independence and vice versa. There is some related work in the context of copulae [26]. Our paper can be seen as a continuation of work by Sarkar [38] and, in particular, by Karlin and Rinott [20, 21]. They noted that the family of MTP_2 distributions is stable with respect to forming marginal and conditional distributions. At least as important is that they give constraints on different types of measures of dependence needed to verify the MTP_2 property of a joint distribution for discrete and for Gaussian random variables.

The MTP_2 property may appear extremely restrictive when higher order interactions are needed to capture the relevant types of conditional dependence or when distributions are studied which do not satisfy any conditional independence constraints. However, as we shall see, the MTP_2 constraints become less restrictive when imposing an additional Markov structure. For example, all finite dimensional distributions of a Markov chain are MTP_2 whenever all 2×2 minors of its transition matrix are nonnegative [20], Proposition 3.10. This result holds true also for nonhomogeneous Markov chains. Moreover, models with latent, that is hidden or unobserved, variables may be MTP_2. For example, factor analysis models with a single factor are MTP_2 when each observed variable has an, albeit unobserved, positive dependence on the single hidden factor [46]. Similar statements apply to binary latent class models [2, 16, 46] and to latent tree models, both in the Gaussian and in the binary setting [40, 50]. Furthermore, many data sets can be well explained or modelled assuming that the generating distribution is MTP_2 or nearly MTP_2; see Section 4 for some examples and also the discussion in Section 8.

The paper is organized as follows: In Section 2, we introduce our notation and provide the main definitions. In Section 3, we review basic properties of MTP_2 distributions and discuss the link to positive dependence and independence structures. In Section 4, we concentrate on the MTP_2 condition in the Gaussian and binary setting and provide several data examples where the MTP_2 property appears naturally. Section 5 analyzes MTP_2 distributions with respect to conditional independence relations. One of the main results in this paper is Theorem 5.3, which shows that any independence model generated by an MTP_2 distribution is a singleton-transitive compositional semi-graphoid which is also upward-stable; the latter means that new arbitrary elements can be added to the conditioning set of every existing independence statement without violating independence. Theorem 5.5 gives a complete characterization of the marginal independence structures of MTP_2 distributions. In Section 6, we study Markov properties of MTP_2 distributions and show that such distributions are always faithful to their concentration graph. In Section 7, we analyze factorization properties of MTP_2 distributions, show how to use these properties to build MTP_2 distributions from smaller MTP_2 distributions, briefly discuss log-linear expansions of discrete MTP_2 distributions, and give conditions for conditional Gaussian distributions to satisfy the MTP_2 constraints. We conclude our paper with a short discussion in Section 8.

2. Preliminaries and notation. Let V be a finite set and let $X = (X_v, v \in V)$ be a random vector. We consider the product space $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, where $\mathcal{X}_v \subseteq \mathbb{R}$ is the state space of X_v, inheriting the order from $\mathbb{R}$. In this paper, the state spaces are either discrete (finite sets) or open intervals on the real line. Hence, we can partition the set of variables as $V = \Delta \cup \Gamma$, where $\mathcal{X}_v$ is discrete if $v \in \Delta$, and $\mathcal{X}_v$ is an open interval if $v \in \Gamma$. For $A \subseteq V$, we further write $X_A = (X_v)_{v \in A}$, $\mathcal{X}_A = \bigtimes_{v \in A} \mathcal{X}_v$ and so on.

All distributions are assumed to have densities with respect to the product measure $\mu = \bigotimes_{v \in V} \mu_v$, where μ_v is the counting measure for $v \in \Delta$, and μ_v is the Lebesgue measure giving length 1 to the unit interval for $v \in \Gamma$. We shall refer to μ as the *standard base measure*. Similarly to the above, we write $\mu_A = \bigotimes_{v \in A} \mu_v$.

Finally, we introduce some definitions related to graphs: an *undirected graph* $G = (V, E)$ (sometimes also called *concentration graph* in the literature of graphical models) consists of a nonempty set of *vertices* or *nodes* V and a set of undirected edges E. Our graphs are *simple* meaning that they have no self-loops and no multiple edges. We write uv for an edge between u and v and say that the vertices u and v are *adjacent*. A *path* in G is a sequence of nodes $(v_0, v_1, \dots, v_k)$ such that $v_i v_{i+1} \in E$ for all $i = 0, \dots, k-1$ and no node is repeated, that is, $v_i \neq v_j$ for all $i, j \in \{0, 1, \dots, k\}$ with $i \neq j$. Thus, an edge is the shortest type of path. A *cycle* is a path with the modification that $v_0 = v_k$. Furthermore, we say that two distinct nodes $u, v \in V$ are *connected* if there is a path between u and v; a graph is *connected* if all pairs of distinct nodes are connected. A graph is *complete* if all possible edges are present. In addition, two subsets $A, B \subseteq V$ are *separated* by $S \subset V \setminus (A \cup B)$ if every path between A and B passes through a node in S. A subgraph of G *induced* by a set $A \subset V$ consists of the nodes in A and of the edges in G between nodes in A. Finally, a maximal complete subgraph is a *clique*.

3. Basic properties and positive dependence. We start this section by formally introducing MTP_2 distributions and discuss basic properties of these. We define the coordinatewise minimum and maximum as

$$x \wedge y = \big(\min(x_v, y_v), v \in V\big), \qquad x \vee y = \big(\max(x_v, y_v), v \in V\big).$$

A function f on $\mathcal{X}$ is said to be *multivariate totally positive of order two* (MTP_2) if

$$f(x)f(y) \leq f(x \wedge y) f(x \vee y) \qquad \text{for all } x, y \in \mathcal{X}. \tag{3.1}$$

For $|V| = 2$, a function that is MTP_2 is simply called *totally positive of order two* (TP_2) [20]. Let $X = (X_v, v \in V)$ have density function f with respect to the standard base measure μ. Then we say that X or the distribution of X is MTP_2 if its density function f is MTP_2. Note that this concept is well defined since $\mathcal{X}_v$ is either discrete or an open interval on the real line.

A basic property of MTP_2 distributions is that it is preserved under increasing coordinatewise transformations. We begin with a simple result for strictly increasing functions.

PROPOSITION 3.1. *Let X be a random vector taking values in $\mathcal{X}$. Let $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathbb{R} \to \mathbb{R}$ are strictly increasing and differentiable for all $v \in V$. If the distribution of X on $\mathcal{X}$ is* MTP_2, *then the distribution of $Y = \phi(X)$ is* MTP_2.

PROOF. We use the following fact from [20], equation (1.13): Let $a_v : \mathbb{R} \to \mathbb{R}$ be positive and let $b_v : \mathbb{R} \to \mathbb{R}$ be nondecreasing. If $f : \mathbb{R}^V \to \mathbb{R}$ is MTP_2, then the function

$$g(y) = f\{b_v(y_v), v \in V\} \prod_{v \in V} a_v(y_v), \tag{3.2}$$

is MTP_2. Let $b_v(y_v) = \phi_v^{-1}(y_v)$ and let $a_v(y_v) = 1/\phi_v'(\phi_v^{-1}(y_v))$, where $\phi_v'(y_v)$ denotes the first derivative of ϕ_v. Then $g(y)$ is the density of $Y = \phi(X)$ and we obtain from (3.2) that Y is MTP_2. $\square$

We say that a function $\phi_v(x) : \mathcal{X}_v \to \mathbb{R}$ is *piecewise constant* if $\phi_v(\mathcal{X}_v)$ is finite and we can then similarly show that the MTP_2 property is preserved under transformations which are piecewise constant and nondecreasing.

PROPOSITION 3.2. *Let X be a random vector taking values in $\mathcal{X}$ as before. For $A \subseteq V$, let $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathcal{X}_v \to \mathbb{R}$ is piecewise constant and nondecreasing for all $v \in A$ and $\phi_v(x_v) = x_v$ for $v \notin A$. If the distribution of X is* MTP_2, *then the distribution of $Y = \phi(X)$ is* MTP_2.

PROOF. If f denotes the density function for X and g the density function for Y, both w.r.t. a standard base measure, we have that

$$g(y) = \int_{\phi_A^{-1}(y_A)} f(x)\, d\mu_A(x_A).$$

Since ϕ is nondecreasing, we have

$$\phi_A^{-1}(y_A^1) \wedge \phi_A^{-1}(y_A^2) = \phi_A^{-1}(y_A^1 \wedge y_A^2), \qquad \phi_A^{-1}(y_A^1) \vee \phi_A^{-1}(y_A^2) = \phi_A^{-1}(y_A^1 \vee y_A^2),$$

where for two sets A, B,

$$A \wedge B = \{a \wedge b | a \in A, b \in B\}, \qquad A \vee B = \{a \vee b | a \in A, b \in B\}.$$

Hence, we can apply [20], Corollary 2.1, to obtain that Y is MTP_2. $\square$

COROLLARY 3.3. *Let X be a random vector taking values in $\mathcal{X}$ as before and $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathcal{X}_v \to \mathbb{R}$ is piecewise constant and nondecreasing for all $v \in A$ and $\phi_v(x_v)$ is strictly increasing and differentiable for $v \notin A$. If the distribution of X is* MTP_2, *then the distribution of $Y = \phi(X)$ is* MTP_2.

PROOF. Just combine Proposition 3.1 and Proposition 3.2 by first transforming to $Z_v = X_v$ for $v \in A$ and $Z_v = \phi_v(X_v)$ for $v \notin A$; then subsequently letting $Y_v = \phi_v(Z_v)$ for $v \in A$ and $Y_v = Z_v$ for $v \notin A$. $\square$

The following result establishes that the MTP_2 property is preserved under conditioning, marginalization and monotone coarsening. A *monotone coarsening* is an operation on a finite discrete state space $\mathcal{X}_i$ that identifies a collection of neighboring (in the given total order) states. For example, if $\mathcal{X}_i = \{i_1, \dots, i_p\}$ then $\mathcal{X}_i' = \{\{i_1, \dots, i_j\}, i_{j+1}, \dots, i_k, \{i_{k+1}, \dots, i_p\}\}$ is a monotone coarsening.

PROPOSITION 3.4. *The* MTP_2 *property is closed under conditioning, marginalization and monotone coarsening. More precisely*:

(i) *If X has an* MTP_2 *distribution, then for every $C \subseteq V$, the conditional distribution of $X_C | X_{V \setminus C} = x_{V \setminus C}$ is* MTP_2 *for almost all $x_{V \setminus C}$.*

(ii) *If X has an* MTP_2 *distribution, then for every $A \subseteq V$, the marginal distribution X_A of X is* MTP_2.

(iii) *If X is* MTP_2 *and discrete*, *and Y is obtained from X by monotone coarsening*, *then Y is* MTP_2.

PROOF. Property (i) follows directly from the definition of MTP_2. Property (ii) is shown in [20], Proposition 3.2. Property (iii) is an instance of a nondecreasing and piecewise constant transformation and follows from Proposition 3.2. □

As we will see in the following, properties (i) and (ii) are the fundamental building blocks for understanding the implications of MTP_2 on Markov properties and vice versa. Property (iii) has direct relevance for applications. In the statistical literature, it is often warned that dependence relations may get distorted when combining neighboring levels of discrete variables; see, for instance, [35]. This may still be true for MTP_2 distributions (see Example 6.2 below), but the coarsening property (iii) implies that associations cannot become negative by such a process.

Another interesting fact about the MTP_2 property is that, under suitable support conditions, it is a pairwise property meaning that it can be checked on the level of two variables only, when the remaining variables are fixed. We say that f has *interval support* if for any $x, y \in \mathcal{X}$ the following holds:

$$f(x)f(y) \neq 0 \quad \text{implies} \quad f(z) \neq 0 \qquad \text{for any } x \wedge y \leq z \leq x \vee y. \tag{3.3}$$

Note that having *interval support* is equivalent to having full support over a restricted state space that is a product of intervals. In this setting, Karlin and Rinott [20] prove the following result in their Proposition 2.1.

PROPOSITION 3.5. *If f has interval support and $f : \mathcal{X} \to \mathbb{R}$ is* TP_2 *in every pair of arguments when the remaining arguments are held constant*, *then f is* MTP_2.

We conjecture that this result holds also under a weaker support condition, namely that the support is *coordinatewise connected* [32], meaning that the connected components of the support can be joined by axis-parallel lines. We now provide such an instance in the binary $2 \times 2 \times 2$ setting.

EXAMPLE 3.6. Consider a binary $2 \times 2 \times 2$ distribution, where the support only misses the entries $(1, 0, 0)$ and $(1, 0, 1)$. In this example, there are only two nontrivial pairwise TP_2 constraints to consider, namely for $x_1 = 0$ and for $x_2 = 1$, that is,

$$p_{000}p_{011} \geq p_{010}p_{001} \quad \text{and} \quad p_{010}p_{111} \geq p_{110}p_{011} \tag{3.4}$$

as the remaining six pairwise inequalities reduce to $0 \geq 0$.

So assume p satisfies (3.4). We need to show that the nine inequalities in (4.1) below are satisfied. Again, six of them are trivial, and two of them are exactly those in (3.4). The remaining inequality,

$$p_{000}p_{111} \geq p_{001}p_{110},$$

follows from multiplying the two inequalities in (3.4) to get

$$(p_{000}p_{011})(p_{010}p_{111}) \geq (p_{010}p_{001})(p_{110}p_{011}),$$

and dividing both sides by $p_{010}p_{011}$. Hence, in this case pairwise TP_2 constraints imply the MTP_2 property even though the distribution does not have interval support.

As mentioned in Section 1, if X is MTP_2, then the variables in X are *associated*, that is,

$$\operatorname{cov}\{f(X), g(X)\} \geq 0 \tag{3.5}$$

for any coordinatewise nondecreasing functions f and g for which the covariance exists. For discrete distributions, this follows by the FKG theorem [14], or more generally, by the four functions theorem by Ahlswede and Daykin [1]; the general case was proved by Sarkar [38]. The following result, first proven by Lebowitz [23], shows that the independence structure for associated vectors is encoded in the covariance matrix; see also [17, 28].

THEOREM 3.7 (Corollary 3, [28]). *If X are associated and $\mathbb{E}|X_v|^2 < \infty$ for all $v \in V$, then X_A is independent of X_B if and only if* $\operatorname{cov}(X_u, X_v) = 0$ *for all $u \in A$ and $v \in B$.*

In Section 5 we study conditional independence models for MTP_2 distributions. Interestingly, we will show in Theorem 5.5 that for MTP_2 distributions a stronger result holds, namely that every MTP_2 random vector can be decomposed into independent components such that within each component all variables are mutually marginally dependent. This means in particular that for MTP_2 distributions, all marginal independence relations also hold when conditioning on further variables; the general version of this property will be termed upward-stability; see Section 5.

4. Examples of Gaussian and binary MTP_2 distributions.

Many examples of MTP_2 random variables are discussed in the literature; see, for example, [20, 21]. In this section, we focus on binary and multivariate Gaussian MTP_2 distributions. Although the MTP_2 property may appear restrictive, we want to suggest that MTP_2 distributions are important in practice and in fact appear in real data sets.

4.1. *Multivariate Gaussian* MTP_2 *distributions.*

Consider a multivariate Gaussian random vector X with mean μ and covariance matrix Σ. Denote by K the inverse of Σ. Then the distribution of X is MTP_2 if and only if K is an M-matrix (see [21]), that is:

(i) $k_{vv} > 0$ for all $v \in V$,
(ii) $k_{uv} \leq 0$ for all $u, v \in V$ with $u \neq v$.

Properties and consequences of M-matrices were studied by Ostrowski [29], who chose the name to honour H. Minkowski, who had considered aspects of such matrices earlier. The connection to multivariate Gaussian distributions was established by Bølviken [5].

In the previous section, we mentioned that if X is MTP_2, then its constituent variables are associated. Therefore, for MTP_2 Gaussian distributions it holds that $\sigma_{uv} \geq 0$ for all $u, v \in V$. More precisely, the covariance matrix has a block diagonal structure and each block has only strictly positive elements; see also Theorem 5.5 below. Note, however, that this condition is necessary but not sufficient for the MTP_2 property.

We now analyze by simulation how restrictive the MTP_2 constraint is for Gaussian distributions. We quantify this by studying the ratio of the volume of all correlation matrices that satisfy the MTP_2 constraint to the volume of all correlation matrices. Since no closed-form formula for these volumes is known, we use a simple Monte Carlo simulation. We uniformly sample correlation matrices using the method suggested by Joe [18], which is implemented in the `R` package `clusterGeneration`. We performed simulations for $|V| = 3, 4, 5$ and we here report how many correlation matrices out of 100,000 samples satisfy the MTP_2 constraint.

|V|	**3**	**4**	**5**
MTP_2	5004	90	0

These simulation results show that the relative volume of MTP_2 Gaussian distributions drops dramatically with increasing $|V|$ when no conditional independences are taken into account. However, the picture changes when imposing conditional independence relations. For example, if $|V| = 3$, then by the above simulations about 5% of all Gaussian distributions correspond to MTP_2 distributions. If $1 \perp\!\!\!\perp 2 | 3$, then by a symmetry argument precisely 25% of such distributions are MTP_2. If, in addition, we impose $1 \perp\!\!\!\perp 3 | 2$—which implies also $1 \perp\!\!\!\perp (3, 2)$—the ratio of MTP_2 distributions increases to 50%. Finally, *all* distributions that are fully independent are MTP_2.

We next discuss a prominent data set consisting of the examination marks of 88 students in five different mathematical subjects. The data were reported in [25] and analyzed, for example, in [12, 15, 47]. The inverse of the sample covariance matrix, together with the corresponding partial correlations $\rho_{uv \cdot V \setminus \{u,v\}}$, are displayed in Table 1. This matrix is very close to being an M-matrix with only one negative partial correlation equal to -0.00001. Furthermore, when fitting reasonable graphical models to the data, all fitted distributions are MTP_2.

4.2. *Binary* MTP_2 *distributions.* Suppose that X is a binary random vector with $\mathcal{X} = \{0, 1\}^{|V|}$ and we denote its distribution by $P = [p_x]$ for $x \in \mathcal{X}$. For example, if $|V| = 3$, binary MTP_2 distributions must satisfy the following nine

TABLE 1
Empirical partial correlations (below the diagonal) and entries of the inverse of the sample covariance matrix ($\times 1000$, on and above the diagonal) for the examination marks in five mathematical subjects

	Mechanics	**Vectors**	**Algebra**	**Analysis**	**Statistics**
Mechanics	5.24	−2.44	−2.74	0.01	−0.14
Vectors	0.33	10.43	−4.71	−0.79	−0.17
Algebra	0.23	0.28	26.95	−7.05	−4.70
Analysis	−0.00	0.08	0.43	9.88	−2.02
Statistics	0.02	0.02	0.36	0.25	6.45

inequalities:

$$\begin{aligned}
&p_{011}p_{000} \geq p_{010}p_{001}, \qquad && p_{101}p_{000} \geq p_{100}p_{001},\\
&p_{110}p_{000} \geq p_{100}p_{010}, && p_{111}p_{100} \geq p_{110}p_{101},\\
&p_{111}p_{010} \geq p_{110}p_{011}, && p_{111}p_{001} \geq p_{101}p_{011},\\
&p_{111}p_{000} \geq p_{100}p_{011}, && p_{111}p_{000} \geq p_{010}p_{101},\\
&p_{111}p_{000} \geq p_{001}p_{110}.
\end{aligned} \tag{4.1}$$

The first two rows correspond to the inequalities $p_{x\wedge y}p_{x\vee y} \geq p_x p_y$ as in (3.1), where x and y differ only in two entries. These inequalities are equivalent to requesting that the six possible conditional log-odds ratios are nonnegative. By Proposition 3.5, the inequalities in the last row are implied by the remaining ones in the case when $P > 0$, or more generally, if P has interval support. For $P > 0$, this can be seen from identities of the form

$$\begin{aligned}
p_{111}p_{000} - p_{010}p_{101} &= \frac{p_{000}}{p_{001}}(p_{111}p_{001} - p_{101}p_{011})\\
&\quad + \frac{p_{101}}{p_{001}}(p_{011}p_{000} - p_{010}p_{001}).
\end{aligned}$$

For positive binary distributions, we can verify MTP_2 for any pair of variables with the remaining variables fixed. In the binary case, this gives a single constraint for any choice of a pair and values for the remaining $|V|-2$ variables, in other words $\binom{|V|}{2}\cdot 2^{|V|-2}$ inequalities. For binary MTP_2 distributions, there is a nice description in terms of log-linear parameters in [4]; see also Corollary 7.7 below.

The MTP_2 hypothesis is rather restrictive in the binary setting when no further conditional independence restrictions are assumed. Note, however, that binary models can become more complex than in the Gaussian case, since log-linear interactions of higher-order than pairwise may be present. In the following, we study the volume of MTP_2 distributions with respect to the volume of the whole probability simplex. Similarly, as in the Gaussian setting, we sample uniformly from the probability simplex. We here report how many samples out of 100,000 satisfy the MTP_2 constraints for $|V| = 3, 4$. Note that already for $|V| = 4$ we did not find a single instance although the volume of the set of MTP_2 distributions is always positive:

|V|	**3**	**4**
MTP_2	2195	0

As in the Gaussian case the relative volume of MTP_2 distributions is higher when imposing additional conditional independence restrictions. By the same symmetry argument as in the Gaussian setting, we obtain that for $|V| = 3$ precisely 25% of all binary distributions satisfying $1 \perp\!\!\!\perp 2|3$ are MTP_2. If, in addition, we have $1 \perp\!\!\!\perp 3|2$ then half of these distributions are MTP_2. Finally, all binary full independence distributions are MTP_2.

This interplay with conditional independence might explain in part why binary MTP_2 distributions do arise in practice. See [46], Section 5, for examples of datasets that are MTP_2 or nearly MTP_2. In the following, we discuss two such examples.

EXAMPLE 4.1. We first consider a dataset on *EPH-gestosis*, collected 40 years ago in a study on "Pregnancy and Child Development" by the German Research Foundation and recently analyzed in [46], Section 5.1. EPH-gestosis represents a disease syndrome for pregnant women. The three symptoms are edema (high body water retention), proteinuria (high amounts of urinary proteins) and hypertension (elevated blood pressure). The observed counts $N = (n_x)$ are

$$\begin{bmatrix} n_{000} & n_{010} & n_{001} & n_{011} \\ n_{100} & n_{110} & n_{101} & n_{111} \end{bmatrix} = \begin{bmatrix} 3299 & 107 & 1012 & 58 \\ 78 & 11 & 65 & 19 \end{bmatrix}.$$

If untreated, EPH-gestosis is a major cause of death of mother and child during birth ([41], page 65). However, treatment of the symptoms prevents negative consequences and the symptoms occur rarely after the first pregnancy.

The observed counts have odds-ratios larger than one for each pair at the fixed level of the third variable, hence the empirical distribution is MTP_2. Equivalently, the sample distribution satisfies all the constraints in (4.1). The three symptoms do not occur more frequently jointly than in pairs and the observed conditional odds-ratios are nearly equal given the third symptom. Possible interpretations are that physicians intervened at the latest when two symptoms occurred and that a single common cause, though unknown and unobserved, may have generated the marginal dependences between the symptom pairs.

EXAMPLE 4.2. Next, we discuss an example with five binary random variables. This is a subset of data from a Polish case-control study on laryngeal cancer [49]. Details on the study design, our selection criteria for cases and controls, and the analysis will be given elsewhere.

In case-control studies, the observations are implicitly obtained conditionally on the values of at least one response variable and on relevant explanatory variables. For such designs, the class of concentration graph models are appropriate for studying dependence structure among the variables.

In this study, we have 185 *laryngeal cancer cases* in urban residential areas (coded 1; 35.7%) and 308 controls, coded 0. Four further binary variables are defined so that 1 indicates the level known to carry the higher cancer risk, namely *heavy vodka drinking* (1 := regularly for 2 or more years; 21.3%), *heavy cigarette smoking* (1:= 30 or more cigarettes per day; 13.8%, and 0 := 6 to 29 cigarettes per day), *age at study entry* (1 := 54 to 65 years; 51.5% and 0 := 46 to 53 years), and *level of formal education* (1 := less than 8 years; 57.8 % and 0 := 8 to 11 years).

A well-fitting log-linear model for these data is determined by the sufficient margins $\{\{1,2\},\{1,3\},\{2,3\},\{1,4,5\}\}$, in other words by permitting log-linear interaction terms only among variable groups that are subsets of these sets. This model yields an overall likelihood-ratio χ^2 of 13.6 with 19 degrees of freedom and corresponds to a concentration graph with cliques: $\{1,2,3\}$ and $\{1,4,5\}$. The corresponding observed and fitted counts are

$$\begin{bmatrix} 00000 & 10000 & 01000 & 11000 \\ 00100 & 10100 & 01100 & 11100 \\ 00010 & 10010 & 01010 & 11010 \\ 00110 & 10110 & 01110 & 11110 \\ 00001 & 10001 & 01001 & 11001 \\ 00101 & 10101 & 01101 & 11101 \\ 00011 & 10011 & 01011 & 11011 \\ 00111 & 10111 & 01111 & 11111 \end{bmatrix} : \begin{bmatrix} 85 & 11 & 5 & 6 \\ 10 & 1 & 1 & 2 \\ 46 & 15 & 3 & 7 \\ 7 & 2 & 2 & 5 \\ 51 & 27 & 7 & 18 \\ 4 & 6 & 1 & 4 \\ 73 & 36 & 5 & 30 \\ 5 & 9 & 3 & 6 \end{bmatrix},$$

$$\begin{bmatrix} 85.88 & 9.87 & 7.30 & 6.35 \\ 9.27 & 1.70 & 1.55 & 2.08 \\ 47.59 & 14.31 & 4.19 & 9.21 \\ 5.32 & 2.47 & 0.89 & 3.02 \\ 51.70 & 27.13 & 4.55 & 17.46 \\ 5.78 & 4.68 & 0.97 & 5.73 \\ 70.57 & 39.96 & 6.22 & 25.72 \\ 7.89 & 6.89 & 1.32 & 8.43 \end{bmatrix}.$$

For pairs within the cliques, the fitted two-way margins must coincide with the observed bivariate tables of counts; here, we report marginal observed and fitted odds-ratios, $\mathrm{or}(I, J)$, and fitted conditional odds-ratios given the remaining variables, $\mathrm{or}(I, J|R)$ (see Table 2).

Because the observed $(3, 5)$ odds-ratio is smaller than 1, and hence the log-odds-ratio is negative, the observed distribution is not MTP_2. In addition, 21 of the observed 80 conditional log-odds ratios, $\mathrm{or}(I, J|R)$, are less than 1. However, in the well-fitting model described above, we have $\mathrm{or}(I, J|R) \geq 1$ for all 80

TABLE 2
Observed and fitted marginal and conditional odds-ratios

Variable pair:	**1, 2**	**1, 3**	**1, 4**	**1, 5**	**2, 3**	**2, 4**	**2, 5**	**3, 4**	**3, 5**	**4, 5**
Observed or(I, J):	7.6	1.9	1.7	3.0	2.3	1.4	2.0	1.3	0.9	2.0
Fitted or(I, J):						1.3	1.6	1.1	1.2	
Fitted or$(I, J\|R)$:	7.3	1.5	2.5	4.4	1.9	1	1	1	1	$*$

$*$ 2.4 for controls and 1.02 for cases.

odds-ratios, so that the fitted distribution is MTP_2. This implies that each possible marginal table—here of two, three or four variables—shows positive or vanishing pairwise dependence for all variable pairs.

From the concentration graph, it follows that prediction of drinking and smoking habits cannot be improved by using information about age or level of formal education for the studied cases or controls and that there is no log-linear interaction involving more than three factors. The set of minimal sufficient tables tells that the only three-factor interaction is in the $\{1, 4, 5\}$-table. From the above change in the conditional odds-ratio for pair $(4, 5)$ from 2.4 to 1.02, it follows that the expected improvement in education for younger—compared to older participants—only shows for controls but not for the cases. In combination with the fact that $\mathrm{or}(1, 5|R) = 4.4$, this implies that level of formal education should be explicitly included in comparisons of results across countries and in future studies on laryngeal cancer.

5. Conditional independence models and total positivity. An *independence model* $\mathcal{J}$ over a finite set V is a set of triples $\langle A, B|C\rangle$ (called *independence statements*), where A, B and C are disjoint subsets of V; C may be empty, and $\langle\varnothing, B|C\rangle$ and $\langle A, \varnothing|C\rangle$ are always included in $\mathcal{J}$. The independence statement $\langle A, B|C\rangle$ is read as "A is independent of B given C". Independence models do not necessarily have a probabilistic interpretation; for a discussion on general independence models; see [42].

An independence model $\mathcal{J}$ over a set V is a *semi-graphoid* if it satisfies the following four properties for disjoint subsets A, B, C and D of V:

(S1) $\langle A, B|C\rangle \in \mathcal{J}$ if and only if $\langle B, A|C\rangle \in \mathcal{J}$ (*symmetry*);
(S2) if $\langle A, B \cup D|C\rangle \in \mathcal{J}$, then $\langle A, B|C\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ (*decomposition*);
(S3) if $\langle A, B \cup D|C\rangle \in \mathcal{J}$, then $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C \cup B\rangle \in \mathcal{J}$ (*weak union*);
(S4) $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ if and only if $\langle A, B \cup D|C\rangle \in \mathcal{J}$ (*contraction*).

A semi-graphoid for which the reverse implication of the weak union property holds is said to be a *graphoid* that is, it also satisfies

(S5) if $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C \cup B\rangle \in \mathcal{J}$ then $\langle A, B \cup D|C\rangle \in \mathcal{J}$ (*intersection*).

Furthermore, a graphoid or semi-graphoid for which the reverse implication of the decomposition property holds is said to be *compositional* that is, it also satisfies

(S6) if $\langle A, B|C\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ then $\langle A, B \cup D|C\rangle \in \mathcal{J}$ (*composition*).

Some independence models have additional properties; below we write singleton sets $\{u\}, \{v\}$ compactly as u, v, etc.

(S7) if $\langle u, v|C\rangle \in \mathcal{J}$ and $\langle u, v|C \cup w\rangle \in \mathcal{J}$, then $\langle u, w|C\rangle \in \mathcal{J}$ or $\langle v, w|C\rangle \in \mathcal{J}$ (*singleton-transitivity*);

(S8) if $\langle A, B|C\rangle \in \mathcal{J}$ and $D \subseteq V \setminus (A \cup B)$, then $\langle A, B|C \cup D\rangle \in \mathcal{J}$ (*upward-stability*).

The properties above are not independent and upward-stability is a very strong property. For example, we have the following simple lemma.

LEMMA 5.1. *Any upward stable semi-graphoid satisfies* (S6) *composition*.

PROOF. If $\langle A, B|C\rangle \in \mathcal{J}$, (S8) yields $\langle A, B|C \cup D\rangle \in \mathcal{J}$; hence from (S4) we get that $\langle A, D|C\rangle \in \mathcal{J}$ implies $\langle A, B \cup D|C\rangle \in \mathcal{J}$, which is (S6). $\square$

A fundamental example of an independence model is induced by separation in an undirected graph $G = (V, E)$, denoted by $\mathcal{J}(G)$:

$$\langle A, B|S\rangle \in \mathcal{J}(\mathcal{G}) \iff S \text{ separates } A \text{ from } B$$

in the sense that all paths between A and B intersect S. The independence model $\mathcal{J}(\mathcal{G})$ satisfies all of the above properties (S1)–(S8).

Consider a set V and associated random variables $X = (X_v)_{v \in V}$. For disjoint subsets A, B and C of V we use the short notation $A \perp\!\!\!\perp B|C$ to denote that X_A is *conditionally independent of* X_B *given* X_C [8, 22], that is, that for any measurable $\Omega \subseteq \mathcal{X}_A$ and P-almost all x_B and x_C,

$$P(X_A \in \Omega|X_B = x_B, X_C = x_C) = P(X_A \in \Omega|X_C = x_C).$$

We can now induce an independence model $\mathcal{J}(P)$ by letting

$$\langle A, B|C\rangle \in \mathcal{J}(P) \quad \text{if and only if} \quad A \perp\!\!\!\perp B|C \qquad \text{w.r.t. } P.$$

Probabilistic independence models are always semi-graphoids [31]. If, for example, P has a strictly positive density f, the induced independence model is always a graphoid; see, for example, Proposition 3.1 in [22]. More generally, if f is continuous, Peters [32] showed that the induced independence model is a graphoid if and only if the support is *coordinatewise connected*, that is, all connected components of the support of the density can be connected by axis-parallel lines. In particular, this applies to the discrete case since any function over a discrete space is continuous. See also [9] for general discussions and [37] for necessary and sufficient conditions under which the intersection property holds in joint Gaussian or binary distributions.

Examples of discrete distributions violating one of (S5), (S6) or (S7) have been given in [43]. We will prove in this section that, under weak assumptions, independence models generated by MTP_2 distributions satisfy all of the properties (S1)–(S8).

First, note that by Proposition 3.4 the MTP_2 property is closed under marginalization and conditioning. Applying this to the conditional distribution of (X_u, X_v) given X_C for all $u, v \in V$ with $u \neq v$, $C \subseteq V \setminus \{u, v\}$, implies that the following conditional covariances must be nonnegative:

$$\operatorname{cov}\{\phi(X_u), \psi(X_v) | X_C\} \geq 0 \qquad \text{a.s.} \tag{5.1}$$

for nondecreasing functions $\phi, \psi : \mathbb{R} \to \mathbb{R}$ for which the covariance exists. Recall that a function of several variables is nondecreasing if it is nondecreasing in each coordinate. The following related result was first proved in [38], Section 3.1 (see also [20], Theorem 4.1).

PROPOSITION 5.2. *Let X be* MTP_2. *Then for any subset $A \subseteq V$ and any nondecreasing function $\varphi : \mathcal{X}_A \to \mathbb{R}$ for which $\mathbb{E}|\varphi(X_A)| < \infty$, the conditional expectation*

$$\mathbb{E}\{\varphi(X_A) | X_{V\setminus A} = x_{V\setminus A}\}$$

is nondecreasing in $x_{V\setminus A}$.

We first show that any induced independence model of an MTP_2 distribution is an upward-stable and singleton-transitive compositional semigraphoid, that is, (S1)–(S4) and (S6)–(S8) all hold.

THEOREM 5.3. *Any independence model $\mathcal{J}(P)$ induced by an* MTP_2 *distribution P is an upward-stable and singleton-transitive compositional semigraphoid.*

PROOF. We first note that any probabilistic independence model is a semigraphoid [30]. Next, we establish upward-stability. For this, it suffices to prove that $u \perp\!\!\!\perp v | C$ implies $u \perp\!\!\!\perp v | C \cup \{w\}$ for all $w \in V \setminus (C \cup \{u, v\})$. Since the MTP_2 property is closed under marginalization, it follows that the marginal distribution of $X_{C\cup\{u,v,w\}}$ is MTP_2. Further, because the MTP_2 property is closed under conditioning, after conditioning on C, it suffices to consider only 3 variables and prove the following statement: If the distribution of $X = (X_1, X_2, X_3)$ is MTP_2, then $1 \perp\!\!\!\perp 2$ implies $1 \perp\!\!\!\perp 2|3$.

Recall that $1 \perp\!\!\!\perp 2$ if and only if $\mathbb{P}(X_1 > a_1, X_2 > a_2) = \mathbb{P}(X_1 > a_1)\mathbb{P}(X_2 > a_2)$ for all $a_1, a_2 \in \mathbb{R}$. Equivalently, defining $Y_i = \mathbb{1}(X_i > a_i)$ this translates into $Y_1 \perp\!\!\!\perp Y_2$, which now must be satisfied for any choice of $a_1, a_2 \in \mathbb{R}$. A similar condition holds for the conditional independence $1 \perp\!\!\!\perp 2|3$, in which case we require that $Y_1 \perp\!\!\!\perp Y_2|X_3$ for all $a_1, a_2 \in \mathbb{R}$.

The advantage of working with the indicator functions is that they define bounded random variables, which implies the existence of moments. By Proposition 3.2, the vector (Y_1, Y_2, X_3) is MTP_2. Independence of X_1 and X_2 in combination with the law of total covariance implies that for every choice of $a_1, a_2 \in \mathbb{R}$

$$\begin{aligned} 0 &= \mathrm{cov}(Y_1, Y_2) \\ &= \mathrm{cov}\big(\mathbb{E}(Y_1|X_3), \mathbb{E}(Y_2|X_3)\big) + \mathbb{E}\big(\mathrm{cov}(Y_1, Y_2|X_3)\big). \end{aligned}$$

By Proposition 5.2, $\mathbb{E}(Y_1|X_3 = x_3)$ and $\mathbb{E}(Y_2|X_3 = x_3)$ are almost everywhere nondecreasing and bounded functions of x_3, and hence their covariance exists and is nonnegative by (3.5) and the fact that univariate random variables are always associated. Moreover, it follows from (5.1) that $\mathrm{cov}(Y_1, Y_2|X_3 = x_3) \geq 0$ for almost all x_3, and thus its expectation is nonnegative. This means that we expressed zero as a sum of two nonnegative terms, and thus both terms must be zero. This implies that $\mathrm{cov}(Y_1, Y_2|X_3 = x_3) = 0$ for almost all x_3. Hence, by Theorem 3.7, we obtain that $Y_1 \perp\!\!\!\perp Y_2|X_3$. Now by varying a_1, a_2 we conclude that $X_1 \perp\!\!\!\perp X_2|X_3$. Having established upward-stability, composition now follows from Lemma 5.1.

We finally prove singleton-transitivity. Using upward-stability this property can be rephrased in a simpler form as

$$1 \perp\!\!\!\perp 2 \quad \Longrightarrow \quad 1 \perp\!\!\!\perp 3 \quad \text{or} \quad 2 \perp\!\!\!\perp 3.$$

So we assume that $1 \perp\!\!\!\perp 2$. By the same argument as in the previous paragraph, $\mathrm{cov}(Y_1, Y_2) = 0$ implies that $\mathrm{cov}(\mathbb{E}(Y_1|X_3), \mathbb{E}(Y_2|X_3)) = 0$. By Theorem 5.2, both $f(X_3) := \mathbb{E}(Y_1|X_3)$ and $g(X_3) := \mathbb{E}(Y_2|X_3)$ are nondecreasing functions. Their covariance is zero and can be rewritten as

$$\mathrm{cov}\big(f(X_3), g(X_3)\big) = \int_{\{x' > x\}} \big(f(x') - f(x)\big)\big(g(x') - g(x)\big)\, d\big(\mu_3(x) \otimes \mu_3(x')\big).$$

Note that $\{x' > x\} \supseteq \{f(x') > f(x)\} \cap \{g(x') > g(x)\}$ and so the integral on the right-hand side, which is equal to zero, is bounded from below by

$$\int_{\{f(x') > f(x)\} \cap \{g(x') > g(x)\}} \big(f(x') - f(x)\big)\big(g(x') - g(x)\big)\, d\big(\mu_3(x) \otimes \mu_3(x')\big),$$

which is strictly positive unless the set $\{f(x') > f(x)\} \cap \{g(x') > g(x)\}$ has measure zero. This set is the set of all pairs (x, x') such that $x' > x$ and $f(x') > f(x)$, $g(x') > g(x)$, and its measure is half the measure of the set of all (x, x') such that $f(x') \neq f(x)$ and $g(x') \neq g(x)$. This measure is zero only if either f or g is constant almost everywhere.

It follows that for every $a_1, a_2 \in \mathbb{R}$ either $\mathbb{E}(Y_1|X_3) = \mathbb{P}(X_1 > a_1)$ or $\mathbb{E}(Y_2|X_3) = \mathbb{P}(X_2 > a_2)$. Let $U \subseteq \mathbb{R}^2$ be the set of all (a_1, a_2) such that the former equality holds and V be the set such that the latter holds. Let π_i denote the projection on the ith coordinate in $\mathbb{R}^2$. We have $U \cup V = \mathbb{R}^2$, and so if $\pi_1(U) \neq \mathbb{R}$, then $\pi_2(V) = \mathbb{R}$. This implies that $\pi_1(U) = \mathbb{R}$ or $\pi_2(V) = \mathbb{R}$. For simplicity, assume that the latter holds but the general argument is the same. If $\pi_2(V) = \mathbb{R}$, then for every $a_2 \in \mathbb{R}$ there exists a_1 such that $\mathbb{E}(Y_2|X_3) = \mathbb{P}(X_2 > a_2)$, or equivalently $\mathbb{P}(X_2 > a_2|X_3) = \mathbb{P}(X_2 > a_2)$. We conclude that $2 \perp\!\!\!\perp 3$. This, up to symmetry, implies that $1 \perp\!\!\!\perp 3$ or $2 \perp\!\!\!\perp 3$. $\square$

We now analyze the intersection property. It is important to note that an MTP_2 independence model is not necessarily a graphoid, as the following simple example shows.

EXAMPLE 5.4. Consider the binary MTP_2 distribution with

$$p_{000} = p_{111} = \frac{1}{2}.$$

Then $1 \perp\!\!\!\perp 2|3$ and $1 \perp\!\!\!\perp 3|2$, but $1 \not\perp\!\!\!\perp (2, 3)$ and, therefore, the intersection property does not hold.

As a consequence of the earlier mentioned result by Peters [32], any MTP_2 distribution with continuous density and coordinatewise connected support is an upward-stable and singleton-transitive compositional graphoid.

We conclude this section with the following property of MTP_2 distributions.

THEOREM 5.5. *Let the distribution of X be* MTP_2 *with none of X_v having a degenerate distribution. Then X can be decomposed into independent components such that within each component all variables are mutually marginally dependent.*

PROOF. As in the previous proof, we define $Y_i = \mathbb{1}(X_i > a_i)$ and by Theorem 3.7 it suffices to prove that the covariance matrix of Y is block diagonal with strictly positive entries in each block. We write $u \sim v$ if the covariance between Y_u and Y_v is nonzero and we show that $u \sim v$ is an equivalence relation, and thus induces a partition of V into independent blocks. It is clear that $u \sim u$ and $u \sim v$ whenever $v \sim u$. It remains to show that $u \sim v$ and $v \sim w$ imply $u \sim w$. But if $u \nsim w$, we have $\sigma_{uw} = 0$ and thus $u \perp\!\!\!\perp w$. Using upward-stability from Theorem 5.3 yields $u \perp\!\!\!\perp w|v$ and singleton-transitivity yields $u \perp\!\!\!\perp v$ or $v \perp\!\!\!\perp w$, which contradicts that $u \sim v$ and $v \sim w$. $\square$

6. Faithfulness and total positivity. In the following, we shall write $A \perp_G B|C$ for the graph separation $\langle A, B|C\rangle \in \mathcal{J}(\mathcal{G})$ and $A \perp\!\!\!\perp B|C$ for the relation $\langle A, B|C\rangle \in \mathcal{J}(\mathcal{P})$ in the independence model generated by P. For a graph $G = (V, E)$, an independence model $\mathcal{J}$ defined over V satisfies the *global Markov property* w.r.t. a graph G, if for disjoint subsets A, B and C of V the following holds:

$$A \perp_G B|C \quad \Longrightarrow \quad \langle A, B|C\rangle \in \mathcal{J}.$$

If $\mathcal{J}(P)$ satisfies the global Markov property w.r.t. a graph G, we also say that P *is Markov* w.r.t. G.

We say that an independence model $\mathcal{J}$ is *probabilistic* if there is a distribution P such that $\mathcal{J} = \mathcal{J}(P)$. We then also say that P is *faithful* to $\mathcal{J}$. If P is faithful to $\mathcal{J}(G)$ for a graph G, then we also say that P is *faithful to* G. Thus, if P is faithful to G it is also Markov w.r.t. G.

In this section, we examine the faithfulness property for MTP_2 distributions. Let P denote a distribution on $\mathcal{X}$. The *pairwise independence graph* of P is the undirected graph $G(P) = (V, E(P))$ with

$$uv \notin E(P) \quad \Longleftrightarrow \quad u \perp\!\!\!\perp v|V \setminus \{u, v\}.$$

A distribution P is said to satisfy the *pairwise Markov property* w.r.t. an undirected graph $G = (V, E)$ if

$$uv \notin E \quad \Longrightarrow \quad u \perp\!\!\!\perp v|V \setminus \{u, v\}.$$

Thus, any distribution P satisfies the pairwise Markov property w.r.t. its pairwise independence graph $G(P)$; indeed, $G(P)$ is the smallest graph that makes P pairwise Markov.

Generally, a distribution may be pairwise Markov w.r.t. a graph without being globally Markov. However, if an MTP_2 distribution P satisfies the coordinatewise connected support condition, in particular if it is strictly positive, and since these are sufficient conditions for the intersection property to hold, then pairwise and global Markov properties are equivalent; see [36]. We prove in the following result that if the pairwise-global equivalence is already established, then P is in fact faithful to $G(P)$.

THEOREM 6.1. *If P is* MTP_2 *and its independence model is a graphoid, then P is faithful to its pairwise independence graph $G(P)$.*

PROOF. By Theorem 3.7 in [22], it follows that P is globally Markov w.r.t. $G(P)$.

To establish faithfulness, we consider disjoint subsets A, B and C so that C does not separate A from B in $G(P)$. We need to show that $A \not\perp\!\!\!\perp B|C$.

First, let $uv \in E$. Then $u \not\perp\!\!\!\perp v|V \setminus \{u, v\}$ and hence by upward-stability as shown in Theorem 5.3, $u \not\perp\!\!\!\perp v|C$ for any $C \subset V \setminus \{u, v\}$.

Since C does not separate A from B, there exists $u \in A$ and $v \in B$ and a path $u = v_1, v_2, \dots, v_r = v$ such that $v_k \notin C$ for all $k = 1, \dots, r$, and $v_k v_{k+1} \in E$ for all $k = 1, \dots, r-1$. By the previous argument, we obtain that $v_k \not\perp\!\!\!\perp v_{k+1} | C$ for all $k = 1, \dots, r-1$. By singleton-transitivity, $v_1 \not\perp\!\!\!\perp v_2 | C$ and $v_2 \not\perp\!\!\!\perp v_3 | C$ imply that $v_1 \not\perp\!\!\!\perp v_3 | C$. Repeating this argument yields $u \not\perp\!\!\!\perp v | C$ and hence $A \not\perp\!\!\!\perp B | C$. □

Note that this was also shown by Slawski and Hein [39] in the case of a Gaussian MTP_2 distribution. In fact, in the Gaussian case it follows readily since conditional covariances between any pair of variables can be obtained through adding (nonnegative) partial correlations along paths in $G(P)$ [19]; see also [45].

Notice that if X has coordinatewise connected support, then Theorem 5.5 is a direct corollary of Theorem 6.1. This is because if a distribution is faithful to an undirected graph, then the statement of Theorem 5.5 obviously holds. However, the latter theorem is still interesting as it covers cases that Theorem 6.1 does not; such as that of Example 5.4.

In Section 3, we postponed showing that the stability of the MTP_2 property under coarsening as established in (iii) of Proposition 3.4 does not imply that coarsening preserves conditional independence relations for MTP_2 distributions. We demonstrate this in the following example.

EXAMPLE 6.2. Consider the trivariate discrete distributions of (I, J, K) where I and K are binary taking values in $\{0, 1\}$ whereas J is ternary with state space $\{0, 1, 2\}$ given as follows:

$$p_{ijk} = \theta_{i|j}\phi_{k|j}\psi_j,$$

where $\psi_j = 1/3$ for all j, $\theta_{1|0} = \phi_{1|0} = 1 - \theta_{0|0} = 1 - \phi_{0|0} = 1/4$, $\theta_{1|1} = \phi_{1|1} = 1 - \theta_{0|1} = 1 - \phi_{0|1} = 1/3$, and $\theta_{1|2} = \phi_{1|2} = 1 - \theta_{0|2} = 1 - \phi_{0|2} = 1/2$. This distribution is easily seen to be MTP_2 which also follows from Proposition 7.1 below. By construction, it also satisfies $I \perp\!\!\!\perp K | J$ so that its concentration graph has edges IJ and JK.

Now define the binary variable L by monotone coarsening of J so that $L = 0$ if $J = 0$ and $L = 1$ if $J \in \{1, 2\}$. Letting q_{ilk} denote the joint distribution of (I, L, K) we get, for example,

$$q_{010} = p_{010} + p_{020} = (\theta_{0|1}\phi_{0|1} + \theta_{0|2}\phi_{0|2})/3 = (4/9 + 1/4)/3 = 25/108,$$

and similarly

$$q_{011} = q_{110} = 17/108, \qquad q_{111} = 13/108$$

so that the odds-ratio between I and J conditional on $\{L = 1\}$ becomes

$$\theta = \frac{q_{010}q_{111}}{q_{110}q_{011}} = \frac{13 \times 25}{17^2} = \frac{325}{289} > 1.$$

Hence, after coarsening, the conditional association between I and J given the third variable changes from absent to positive. Note that the MTP_2 property ensures nonnegativity of the distorted association.

Clearly, the distribution after coarsening remains faithful to its concentration graph, but coarsening changes the latter to become the complete graph on $V = \{I, L, K\}$.

For completeness of Theorem 6.1, it is important to show that any concentration graph is realizable by an MTP_2 distribution. We prove this fact in the special case of Gaussian distributions.

PROPOSITION 6.3. *Any undirected graph G is realizable as the concentration graph $G(P)$ of some* MTP_2 *Gaussian distribution.*

PROOF. Let A be the adjacency matrix of G, that is, $A_{ij} = 1$ if and only if (i, j) is an edge of G. Because G is undirected, A is symmetric. Since the set of positive definite matrices is open and the identity matrix is positive definite, it follows that $K = I - \varepsilon A$ is positive definite if $\varepsilon > 0$ is sufficiently small. By construction, K is an M-matrix and its nonzero elements correspond to the edges of the graph G. $\square$

7. Special instances of total positivity. We conclude this paper with a section on how to construct MTP_2 distributions from a collection of smaller MTP_2 distributions, a brief discussion of conditions for the MTP_2 property of discrete distributions in terms of log-linear interaction parameters, and characterizing conditional Gaussian distributions which are MTP_2.

7.1. *Singleton separators.* Let $A, B \subset V$. We then say that two random variables X_A and X_B with distributions P_A and P_B are *consistent* if the distribution of $X_{A\cap B}$ is the same under P_A as under P_B. Then one can define a new distribution denoted by $P_A \star P_B$ and known as the *Markov combination* of P_A with density f and P_B with density g (see [10]). Its density is denoted by $f \star g$ and given by

$$(f \star g)(x_{A\cup B}) = \frac{f(x_A)g(x_B)}{h(x_{A\cap B})}.$$

Here, h denotes the density of $X_{A\cap B}$, common to P_A and P_B. In the following, we show that the Markov combination of two MTP_2 distributions is again MTP_2 as long as they are glued together over a 1-dimensional margin.

PROPOSITION 7.1. *Suppose that $|A \cap B| = 1$. Then the Markov combination $P_A \star P_B$ of a consistent pair of distributions P_A and P_B is* MTP_2 *if and only if P_A and P_B are both* MTP_2.

PROOF. Since P_A and P_B are marginal distributions of $P_A \star P_B$ and the MTP_2 condition is preserved under marginalization, we only need to prove one direction. Assume that P_A and P_B are MTP_2. The product of MTP_2 functions is an MTP_2 function; see, for example, Proposition 3.3 in [20] for this basic result. This implies that $(fg)(x) = f(x_A)g(x_B)$ is MTP_2. Now, if $A \cap B$ is a singleton, then also $f \star g$ is MTP_2 because multiplying by functions of a single variable preserves the MTP_2 property; cf. (3.2). $\square$

For example, Proposition 7.1 implies that for the fitted model in Example 4.2 we only need to check the MTP_2 condition in each of the two clique marginals $\{1, 2, 3\}$ and $\{1, 4, 5\}$ to verify that the fitted distribution is MTP_2. Since the fitted distribution is positive, this involves only $6 + 6 = 12$ log-odds-ratios; see the discussion of (4.1) in Section 4.2. In addition, as there are only pairwise interactions in the $\{1, 2, 3\}$-marginal, conditional log-odds-ratios for any pair of these variables are constant in the third variable, and hence we actually only need to check $3 + 6 = 9$ such ratios to verify the MTP_2 property for the model fitted to the laryngeal cancer data; see also Theorem 7.5 below.

Unfortunately, the conclusion in Proposition 7.1 does not hold in general if $|A \cap B| > 1$, as we show in the following example.

EXAMPLE 7.2. Suppose that $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$, and let $X = (X_1, X_2, X_3, X_4) \in \{0, 1\}^4$. Consider the following distribution:

$$\begin{aligned}
&[p_{0000}, p_{0001}, p_{0010}, p_{0011}, p_{0100}, p_{0101}, p_{0110}, p_{0111}] \\
&\qquad = [1, 2, 2, 20, 2, 20, 20, 400]/Z, \\
&[p_{1000}, p_{1001}, p_{1010}, p_{1011}, p_{1100}, p_{1101}, p_{1110}, p_{1111}] \\
&\qquad = [2, 4, 20, 200, 20, 200, 400, 8000]/Z,
\end{aligned}$$

where the normalizing constant $Z = 9313$. It is easy to check that for every $i, j, k, l \in \{0, 1\}$ the following holds:

$$p_{ijkl} = \frac{p_{ijk+}p_{+jkl}}{p_{+jk+}}.$$

Hence, the distribution $P = [p_{ijkl}]$ can be obtained as the Markov combination of two distributions, namely p_{ijk+} over $\{1, 2, 3\}$ and p_{+jkl} over $\{2, 3, 4\}$. One can also easily check that both these distributions are MTP_2. However, since

$$\begin{aligned}
&p_{(1,1,0,1)\wedge(1,0,1,1)}\, p_{(1,1,0,1)\vee(1,0,1,1)} - p_{(1,1,0,1)}\, p_{(1,0,1,1)} \\
&\qquad = p_{1001}p_{1111} - p_{1101}p_{1011} = -\frac{8000}{9313^2},
\end{aligned}$$

P is not MTP_2.

As a direct consequence of Proposition 7.1, we obtain the following result for *decomposable graphs*, which are graphs where there is no cycle of length more than three such that all its nonneighboring nodes (on the cycle) are not adjacent (see, e.g., [22] for a review).

COROLLARY 7.3. *Let G be a decomposable graph such that the intersection of any two cliques is either empty or a singleton. Let P be a distribution that is Markov w.r.t. G. Then P is* MTP_2 *if and only if the marginal distribution over each clique is* MTP_2.

PROOF. The proof follows by induction over the number of cliques. □

As we show in the following example, Corollary 7.3 cannot be extended directly to nondecomposable graphs. It does not hold in general that a distribution is MTP_2 if the margins over all cliques in the graph are MTP_2 and the cliques intersect in singletons only. However, as we show in Theorem 7.5 such a result does hold if all clique potentials are MTP_2 functions.

EXAMPLE 7.4. Consider the following 4-dimensional binary distribution $P = [p_{ijkl}]$ with

$$p_{ijkl} = \frac{1}{Z} A_{ij} B_{jk} C_{kl} D_{il},$$

where

$$Z = 243, \qquad A = \begin{bmatrix} 6 & 5 \\ 4 & 3 \end{bmatrix} \quad \text{and} \quad B = C = D = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

This distribution is Markov w.r.t. the 4-cycle. We now show that the marginal distributions over each edge are MTP_2. For this, note that a binary 2-dimensional random vector is MTP_2 if and only if its covariance is nonnegative. To see this, observe that in the bivariate case there is only one inequality $p_{00}p_{11} - p_{01}p_{10} \geq 0$. Using the fact that $p_{00} + p_{01} + p_{10} + p_{11} = 1$, it is seen that this inequality is equivalent to $p_{11} - p_{1+}p_{+1} \geq 0$, but $\mathrm{cov}(X_1, X_2) = p_{11} - p_{1+}p_{+1}$.

In this example,

$$\mathrm{cov}(X_1, X_2) = \frac{148}{243^2}, \qquad \mathrm{cov}(X_2, X_3) = \frac{4812}{243^2},$$
$$\mathrm{cov}(X_3, X_4) = \frac{4842}{243^2}, \qquad \mathrm{cov}(X_1, X_4) = \frac{4632}{243^2},$$

and hence all edge-marginals are MTP_2. However, the full distribution P is not MTP_2, since

$$p_{0011}p_{1111} - p_{0111}p_{1011} = -\frac{32}{243^2},$$

which completes the proof by a similar argument as in Example 7.2.

We now show how to overcome these limitations and build MTP_2 distributions over nondecomposable graphs, namely by using MTP_2 potentials over the edges instead of MTP_2 marginal distributions over the edges.

THEOREM 7.5. *A distribution of the form*

$$p(x) = \frac{1}{Z} \prod_{uv \in E} \psi_{uv}(x_u, x_v),$$

where ψ_{uv} are positive functions and Z is a normalizing constant, is MTP_2 *if and only if each ψ_{uv} is an* MTP_2 *function.*

PROOF. Since the distribution p is strictly positive, by Proposition 3.5 p satisfies MTP_2 if and only if it does so for $x, y \in \mathcal{X}$ that differ in two coordinates, say with indices u, v. Write E_u for the set of edges that contain u but not v and E_v for the set of edges that contain v but not u. First, consider the case where $uv \in E$. Then we have that $p(x \wedge y)p(x \vee y) - p(x)p(y) \geq 0$ if and only if

$$\begin{aligned}\psi_{uv}\bigl((x \wedge y)_{uv}\bigr)\psi_{uv}\bigl((x \vee y)_{uv}\bigr) &\prod_{st \in E_u \cup E_v} \psi_{st}\bigl((x \wedge y)_{st}\bigr)\psi_{st}\bigl((x \vee y)_{st}\bigr) \\ \geq \psi_{uv}(x_{uv})\psi_{uv}(y_{uv}) &\prod_{st \in E_u \cup E_v} \psi_{st}(x_{st})\psi_{st}(y_{st}).\end{aligned}$$

All other terms cancel because of the assumption that $x_w = y_w$ for $w \in V \setminus \{u, v\}$. Now note that for $st \in E_u \cup E_v$ we have $\{x_{st}, y_{st}\} = \{(x \wedge y)_{st}, (x \vee y)_{st}\}$ and so the above inequality holds if and only if

$$\psi_{uv}\bigl((x \wedge y)_{uv}\bigr)\psi_{uv}\bigl((x \vee y)_{uv}\bigr) \geq \psi_{uv}(x_{uv})\psi_{uv}(y_{uv}),$$

that is, if and only if ψ_{uv} is MTP_2.

Next, consider the case where $uv \notin E$. By the same argument, one concludes that in this case the above inequalities are in fact equalities, which completes the proof. □

As a final remark note that Theorem 7.5 can directly be extended to distributions of the form

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C),$$

where $\mathcal{C}$ is a family of subsets of V such that for any two $C, C' \in \mathcal{C}$ we have $|C \cap C'| \in \{0, 1\}$.

7.2. *Log-linear interactions*. We next give a short discussion of interaction representations for discrete MTP_2 distributions, as they typically are used in log-linear models for contingency tables. Suppose that $X = (X_v)_{v \in V}$ is a random vector with values in $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$ where each $\mathcal{X}_v$ is finite. Let $\mathcal{D}$ denote the set of subsets of V. Any function $h : \mathcal{X} \to \mathbb{R}^n$ of $\mathcal{X}$ can be expanded as

$$h(x) = \sum_{D \in \mathcal{D}} \theta_D(x), \tag{7.1}$$

where θ_D are functions on $\mathcal{X}$ that only depend on x through x_D, that is, θ_D satisfy that $\theta_D(x) = \theta_D(x_D)$. In the case where $h(x) = \log p(x)$ where p is a positive probability distributions over $\mathcal{X}$, the functions $\theta_D(x)$ are known as *interactions* among variables in D and we shall also use this expression for a general function h.

Without loss of generality, we may assume that $\min \mathcal{X}_v = 0$ for all $v \in V$ and to assure that the representation is unique, we may require that $\theta_D(x) = 0$ whenever $x_d = 0$ for some $d \in D$. With this convention, the sum in (7.1) can be rewritten so it only extends over such $D \in \mathcal{D}$ which are contained in the *support* $S(x)$ of x where $d \in S(x) \iff x_d \neq 0$. In the binary case, when $d_v = 1$, this allows us to use a simpler notation, namely $\theta_D(\mathbf{1}_D) := \theta_D$ for all $D \in \mathcal{D}$.

For any such interaction expansion and a fixed pair $u, w \in V$, we define a function γ_{uw} on $\mathcal{X}$ by

$$\gamma_{uw}(x) = \sum_{D : \{u,w\} \subseteq D \subseteq S(x)} \theta_D(x).$$

Observe that then $\gamma_{uw}(x) = 0$ unless $u, w \in S(x)$, and thus in particular whenever $|S(x)| \leq 1$; further, γ_{uw} is a linear combination of interaction terms.

Let $\mathcal{Z} \subseteq \mathcal{X}$ be a subset of $\mathcal{X}$ which is closed under $\wedge$ and $\vee$. A function $g : \mathcal{Z} \to \mathbb{R}$ is *supermodular* if

$$g(x \wedge y) + g(x \vee y) \geq g(x) + g(y) \qquad \text{for all } x, y \in \mathcal{Z}.$$

Thus, g is supermodular on $\mathcal{X}$ if and only if $\exp(g)$ is MTP_2 on $\mathcal{X}$. A function g is *modular* if both of g and $-g$ are supermodular.

Denote by $\mathcal{X}^A$ the set of all $x \in \mathcal{X}$ with $S(x) = A$. Clearly, $\mathcal{X}^A$ is closed under $\wedge$ and $\vee$. Then we obtain the following result.

THEOREM 7.6. *Let P be a strictly positive distribution of X. Then P is* MTP_2 *if and only if for all $A \subseteq V$ with $|A| \geq 2$ and any given $u, w \in V$ the function γ_{uw} is nonnegative, nondecreasing and supermodular over $\mathcal{X}^A$.*

PROOF. By Proposition 3.5, p is MTP_2 if and only if

$$\log p(x \wedge y) + \log p(x \vee y) - \log p(x) - \log p(y) \geq 0 \tag{7.2}$$

for all $x, y \in \mathcal{X}$ that differ only in two entries. Let $u, w \in V$ and take $x, y \in \mathcal{X}$ satisfying $x_v = y_v$ for all $v \in V \setminus \{u, w\}$. Without loss of generality, we can assume

$x_u < y_u$ and $y_w < x_w$ for otherwise the inequality is trivially satisfied. Using the expansion (7.1), the inequality (7.2) becomes

$$\sum_{D\in\mathcal{D}} \big(\theta_D\big((x\wedge y)_D\big) + \theta_D\big((x\vee y)_D\big) - \theta_D(x_D) - \theta_D(y_D)\big) \ge 0. \tag{7.3}$$

For every $D \subseteq V \setminus \{w\}$, we have $(x \wedge y)_D = x_D$ and $(x \vee y)_D = y_D$. Similarly, for every $D \subseteq V \setminus \{u\}$ we have $(x \wedge y)_D = y_D$ and $(x \vee y)_D = x_D$, and thus, in both cases the corresponding summands in (7.3) are zero. It follows that (7.3) is equivalent to

$$\sum_{D:\{u,w\}\subseteq D\subseteq A} \big(\theta_D\big((x\wedge y)_D\big) + \theta_D\big((x\vee y)_D\big) - \theta_D(x_D) - \theta_D(y_D)\big) \ge 0, \tag{7.4}$$

where $A \subseteq V$ is the support of $x \vee y$ (if D is not contained in A all terms θ_D are zero by our convention).

We now show that the fact that (7.4) must hold for all $u, w \in V$ and $x, y \in \mathcal{X}$ as above is equivalent to the fact that all γ_{uw} satisfy the conditions of the theorem.

Consider three possible cases:

(a) $S(x) = A \setminus \{u\}$, $S(y) = A \setminus \{w\}$,
(b) either $S(x) = A \setminus \{u\}$, $S(y) = A$ or $S(x) = A$, $S(y) = A \setminus \{w\}$;
(c) $S(x) = S(y) = A$.

In other words: in case (a), we have $x_u = y_w = 0$; in case (b), either $x_u = 0$, $y_w > 0$ or $x_u > 0$, $y_w = 0$; and in case (c) we have $x_u, y_w > 0$.

In case (a), we have $\theta_D((x \wedge y)_D) = \theta_D(x_D) = \theta_D(y_D) = 0$ for every D containing $\{u, w\}$ so (7.4) becomes

$$\sum_{D:\{u,w\}\subseteq D\subseteq A} \theta_D\big((x\vee y)_D\big) \ge 0.$$

By choosing different pairs x, y, this can be equivalently rewritten as

$$\sum_{D:\{u,w\}\subseteq D\subseteq S(x)} \theta_D(x_D) \ge 0 \qquad \text{for all } x \in \mathcal{X}$$

and the sum on the left is precisely $\gamma_{uw}(x)$.

In case (b), if $S(x) = A \setminus \{u\}$, $S(y) = A$, (7.4) becomes

$$\sum_{D:\{u,w\}\subseteq D\subseteq A} \big(\theta_D\big((x\vee y)_D\big) - \theta_D(y_D)\big) \ge 0,$$

where $A = S(x \vee y) = S(y)$, which is equivalent to γ_{uw} being nondecreasing on $\mathcal{X}^A$.

Finally, in case (c), all $x \vee y$, $x \wedge y$, x, y have the same support. Thus, γ_{uw} must be supermodular over each $\mathcal{X}^A$. □

As a special case, we recover the characterization of binary MTP_2 distributions in [4].

COROLLARY 7.7. *Let P be a binary distribution with*

$$\log p(x) = \sum_{D:D\subseteq S(x)} \theta_D$$

using the convention that $\theta_D = \theta_D(\mathbf{1}_D)$. Then P is MTP$_2$ *if and only if for all A with $|A| \geq 2$ and all $\{u, w\} \subseteq V$ we have*

$$\sum_{D:\{u,w\}\subseteq D\subseteq A} \theta_D \geq 0.$$

PROOF. In the binary case, each $\mathcal{X}^A$ has only one element and so the only constraint from Theorem 7.6 is the nonnegativity constraint. $\square$

EXAMPLE 7.8. Let $X = (X_1 = 1_A, X_2 = 1_B, X_3 = 1_C)$ be the vector of binary indicator functions of events A, B, C. Reichenbach [34], page 190 (using a different notation) says that *an event B is causally between A and C* if $P(C|B \wedge A) = P(C|B)$ and further

$$1 > P(C|B) > P(C|A) > P(C) > 0,$$
$$1 > P(A|B) > P(A|C) > P(A) > 0.$$

Equivalently, as defined in [6], B is causally between A and C if the following hold:

$$P(A \wedge C) > P(A)P(C), \tag{7.5}$$
$$P(A|B) > P(A|C), \tag{7.6}$$
$$P(C|B) > P(C|A), \tag{7.7}$$
$$P(A \wedge C|B) = P(A|B)P(C|B), \tag{7.8}$$
$$P(\neg A \wedge B) > 0, \qquad P(\neg C \wedge B) > 0. \tag{7.9}$$

In general, causal betweenness does not imply MTP$_2$; if we let $p_{101} = 0$, $p_{000} = 4/10$, and $p_{ijk} = 1/10$ for the remaining six possibilities, B is causally between A and C, but X is not MTP$_2$ since $0 = p_{101}p_{000} < p_{100}p_{001}$.

However, *if $P(X = x) > 0$ for all x and B is causally between A and C, then P is* MTP$_2$. To see this, we expand P in log-linear interaction parameters to get

$$1_A \perp\!\!\!\perp 1_C|(1_B = 1) \quad \Longleftrightarrow \quad \theta_{AC} + \theta_{ABC} = 0. \tag{7.10}$$

Further, a simple but somewhat tedious calculation using (7.10) yields that (7.6) and (7.7) hold if and only if

$$\theta_{AB} > \theta_{AC}, \qquad \theta_{BC} > \theta_{AC},$$

which in combination with (7.10) gives that (7.6)–(7.8) hold if and only if

$$\theta_{AB} + \theta_{ABC} > 0, \qquad \theta_{BC} + \theta_{ABC} > 0, \qquad \theta_{AC} + \theta_{ABC} = 0. \tag{7.11}$$

Thus, Corollary 7.7 ensures that P is MTP_2.

Conversely, if $P(X = x) > 0$ for all x and P is MTP_2, the condition $1_A \perp\!\!\!\perp 1_C \mid (1_B = 1)$ implies "weak causal betweenness", that is, P satisfies the inequalities (7.5)–(7.7) with $>$ replaced by $\geq$; this is true because the weak form of (7.6) and (7.7) follows from the weak form of (7.11), and (7.5) in its weak form expresses that $\mathrm{cov}(X_1, X_3) \geq 0$, which is also a consequence of MTP_2.

Finally, if $P(X = x) > 0$ for all x, P is MTP_2, and the independence graph of P is 1—2—3, then B is causally between A and C, as then P is faithful by Theorem 6.1, which ensures that inequalities are strict.

7.3. *Conditional Gaussian distributions.* In this section, we study *CG-distributions* satisfying the MTP_2 property. The density of a CG-distribution is given by specifying a strictly positive distribution $p(i)$ over the discrete variables for $i \in \mathcal{X}_\Delta$. Then the joint density $f(x) = f(i, y)$ is determined by specifying $f(y|i)$ to be the density of a Gaussian distribution $\mathcal{N}_\Gamma(\xi(i), \Sigma(i))$, where $\xi(i) \in \mathbb{R}^\Gamma$ is the mean vector and $\Sigma(i)$ is the covariance matrix. CG-distributions can also be represented by the set of canonical characteristics (g, h, K) where

$$\log f(x) = \log f(y, i) = g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y;$$

see [22]. Here, $K(i) = \Sigma^{-1}(i)$ is the conditional concentration matrix. We shall say that a function $u(i)$ is *additive* if it has the form

$$u(i) = \sum_{\delta \in \Delta} \alpha_\delta(i_\delta).$$

Before we characterize CG-distribution with the MTP_2 property, we need a small lemma.

LEMMA 7.9. *A function $u : \mathcal{X}_\Delta \to \mathbb{R}$ is additive if and only if it is modular.*

PROOF. If u is additive, then it is clearly modular. To show the converse, we make a log-linear expansion of u as in (7.1)

$$u(i) = \sum_{D \in \mathcal{D}} \eta_D(i).$$

We shall show that if u is modular, then $\eta_D(i) = 0$ whenever $|D| \geq 2$. If for $C \subseteq V$, we let

$$w_C(i) = u(i_C, 0_{V \setminus C}),$$

it follows from the Möbius inversion lemma, also known as inclusion–exclusion (for example p. 239 of [22]) that

$$\eta_D(i) = \sum_{A : A \subseteq D} (-1)^{|D \setminus A|} w_A(i).$$

If $|D| \geq 2$, we can for distinct $u, v \in D$ rewrite this as

$$\eta_D(i) = \sum_{A: A \subseteq D \setminus \{u,v\}} (-1)^{|D \setminus A|} \{w_{A \cup \{u,v\}}(i) - w_{A \cup \{u\}}(i) - w_{A \cup \{v\}}(i) + w_A(i)\}.$$

If u is modular, all terms inside the curly brackets are zero, and hence u is additive. □

PROPOSITION 7.10. *A CG-distribution P with canonical characteristics (g, h, K) is* MTP_2 *if and only if*:

(i) $g(i)$ *is supermodular*;
(ii) $h(i)$ *is additive and nondecreasing*;
(iii) $K(i) = K$ *for all i where K is an M-matrix.*

PROOF. By Proposition 3.5, a CG-distribution is MTP_2 if and only if it satisfies

$$f(y \wedge z, i \wedge j) f(y \vee z, i \vee j) \geq f(y, i) f(z, j)$$

for cases where (y, i) and (z, j) differ on two coordinates. Suppose first that $i = j$ and y, z differ on two coordinates. Then we equivalently need to check whether

$$f(y \wedge z|i) f(y \vee z|i) \geq f(y|i) f(z|i).$$

Since $f(y|i)$ is the density of a Gaussian distribution, this inequality holds for every $y, z \in \mathbb{R}^\Gamma$ and i if and only if each $K(i)$ is an M-matrix.

If i, j and y, z both differ on one coordinate then without loss of generality we can assume $i < j$ and $y > z$ so that $i = i \wedge j$ and $z = y \wedge z$. In this case, we need to show that

$$\log f(z, i) + \log f(y, j) \geq \log f(y, i) + \log f(z, j). \tag{7.12}$$

Write $y = z + te_k$ for some $t > 0$, where e_k is a unit vector in $\mathbb{R}^\Gamma$. Then equivalently

$$\frac{1}{t}\big(\log f(z + te_k, j) - \log f(z, j)\big) \geq \frac{1}{t}\big(\log f(z + te_k, i) - \log f(z, i)\big).$$

Since this holds for every $t > 0$, we can take the limit $t \to 0$, which implies that necessarily

$$\nabla_z \log f(z, j) \geq \nabla_z \log f(z, i) \qquad \text{for all } z \in \mathbb{R}^\Gamma, i < j \in \Delta.$$

Since $\nabla_y \log f(y, i) = h(i) - K(i)y$, this is equivalent to

$$h(j) - h(i) - \big(K(j) - K(i)\big)z \geq 0.$$

The function on the left-hand side is linear in z, and thus this holds for every z if and only if $K(j) = K(i)$ for every i, j and $h(i)$ is nondecreasing in i.

If $y = z$ and i, j differ on two coordinates, using all the conditions that have been already proven to be necessary we need to check that

$$\begin{aligned}(7.13)\qquad &\big(g(i \wedge j) + g(i \vee j) - g(i) - g(j)\big) \\ &\quad + \big(h(i \wedge j) + h(i \vee j) - h(i) - h(j)\big)^T z \geq 0.\end{aligned}$$

This can hold for every z only if h is modular, that is,

$$(7.14)\qquad h(i \wedge j) + h(i \vee j) - h(i) - h(j) = 0 \qquad \text{for all } i, j.$$

Now if (7.14) holds, (7.13) holds if and only if $g(i)$ is super-modular. By Lemma 7.9, h is additive, which completes the proof. $\square$

Proposition 7.10 gives a simple condition for CG-distributions to be MTP_2 in terms of their canonical characteristics. This also implies that the *moment characteristics* (p, ξ, Σ) have simple properties.

PROPOSITION 7.11. *If a CG-distribution is* MTP_2, *its moment characteristics* (p, ξ, Σ) *satisfy*:

(i) $p(i)$ *is* MTP_2;
(ii) $\xi(i)$ *is additive and nondecreasing*;
(iii) $\Sigma(i) = \Sigma$ *for all* i *and all elements of* Σ *are nonnegative.*

PROOF. If the CG-distribution is MTP_2, (iii) follows directly from Proposition 7.10. The condition (i) follows since marginals of MTP_2 distributions are MTP_2 and (ii) follows from (ii) of Proposition 7.10 since $\xi(i) = \Sigma h(i)$ and Σ has only nonnegative elements. $\square$

Thus, MTP_2 CG-distributions are in particular *homogeneous*—$\Sigma(i)$ constant in i—and *mean-additive* [12, 22]. Note that the converse of Proposition 7.11 is not true since $\xi(i)$ can be nonincreasing and $h = K\xi(i)$ decreasing, even when K is an M-matrix.

Finally, we make expansions of $g(i)$ as in Section 7.2:

$$g(i) = \sum_{D \in \mathcal{D}} \lambda_D(i), \qquad \gamma^g_{uw}(i) = \sum_{D:\{u,w\} \subseteq D \subseteq S(i)} \lambda_D(i)$$

and recall that $\mathcal{I}^A = \{i : S(i) = A\}$. We then have the following alternative formulation of Proposition 7.10.

COROLLARY 7.12. *A CG-distribution is* MTP_2 *if and only if*:

(i) *For all* $A \subseteq V$ *with* $|A| \geq 2$ *and all* $u, w \in V$, *the functions* $\gamma^g_{uw}(i)$ *are supermodular and nondecreasing over each* $\mathcal{I}^A$;

(ii) *The function $h(i)$ is additive*

$$h(i) = \sum_{\delta \in \Delta} \alpha_\delta(i_\delta)$$

with nondecreasing components $\alpha_\delta(i_\delta)_v$;

(iii) *There exists an M-matrix K such that $K(i) = K$ for all i.*

PROOF. The proof of Theorem 7.6 does not use that $\log p$ is the logarithm of a probability distribution; hence the corresponding conclusions also apply to the expansion of g. □

Note that if i is binary, condition (i) of the Corollary simplifies as in Corollary 7.7 to the condition that for all $A \subseteq V$ and all $u \neq w \in A$ we have

$$\sum_{D:\{u,w\} \subseteq D \subseteq A} \lambda_D \geq 0.$$

8. Discussion. In this paper, we showed that MTP_2 distributions enjoy many important properties related to conditional independence; in particular, an independence model generated by an MTP_2 distribution is an upward-stable and singleton-transitive compositional semi-graphoid which is faithful to its concentration graph if it has coordinatewise connected support.

We illustrated with several examples that MTP_2 models are useful for data analysis. The MTP_2 constraint seems restrictive when no conditional independences are taken into account. However, the picture changes and the MTP_2 constraint becomes less restrictive when imposing conditional independence constraints in the form of Markov properties.

An important property of MTP_2 models, which is of practical relevance, is that the positive conditional dependence of two variables given all remaining variables implies a positive dependence given any subset of the remaining variables. This is a highly desirable feature, especially when results of follow-up empirical studies are to be compared with an earlier comprehensive study, and the studies coincide only in a subset of core variables.

More generally, in any MTP_2 concentration graph model, dependence reversals cannot occur. This undesirable, worrying feature has been described and studied under different names depending on the types of the involved variables, for instance as *near multicollinearity* for continuous variables, as the *Yule–Simpson paradox* for discrete variables, or as the effects of *highly unbalanced experimental designs* with explanatory discrete variables for continuous responses. It is a remarkable feature of MTP_2 distributions that such dependence reversals are absent.

These observations suggest that it would be desirable to develop further methods for hypothesis testing and estimation under the MTP_2 constraint as done for the binary case in [4]. Our results may also be applied not to the joint distribution of all variables, but only to the joint distribution of a subset of the variables, given a fixed set of level combinations of the remaining variables. This is particularly interesting in empirical studies, where a set of possible regressors and background variables is manipulated to make the studied groups of individuals as comparable as possible; for instance, by selecting equal numbers of persons for fixed level combinations of some features, by proportional allocation of patients to treatments, by matching or by stratified sampling. In such situations, not much can be inferred from the study results about the conditional distribution of the manipulated variables given the responses. However, the MTP_2 property of the joint conditional distribution of the responses given the manipulated variables could be essential.

Acknowledgments. This article represents research initiated at the American Institute of Mathematics workshop "Positivity, Graphical Models, and the Modeling of Complex Multivariate Dependencies" in October 2014. The authors thank the AIM and NSF for their support and are indebted to other participants of the workshop for useful and constructive discussions. We are especially grateful to Hélène Massam for her constructive remarks and encouragement. Thanks are also due to Heiko Becher for permission to use his data in Example 4.2, and to Antonio Forcina and two anonymous referees for their helpful comments.

REFERENCES

[1] AHLSWEDE, R. and DAYKIN, D. E. (1978). An inequality for the weights of two families of sets, their unions and intersections. *Z. Wahrsch. Verw. Gebiete* **43** 183–185. MR0491189

[2] ALLMAN, E. S., RHODES, J. A., STURMFELS, B. and ZWIERNIK, P. (2015). Tensors of nonnegative rank two. *Linear Algebra Appl.* **473** 37–53. MR3338324

[3] BARTOLUCCI, F. and BESAG, J. (2002). A recursive algorithm for Markov random fields. *Biometrika* **89** 724–730. MR1929176

[4] BARTOLUCCI, F. and FORCINA, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *Ann. Statist.* **28** 1206–1218. MR1811325

[5] BØLVIKEN, E. (1982). Probability inequalities for the multivariate normal with nonnegative partial correlations. *Scand. J. Stat.* **9** 49–58. MR0651862

[6] CHVÁTAL, V. and WU, B. (2011). On Reichenbach's causal betweenness. *Erkenntnis* **76** 41–48.

[7] COLANGELO, A., SCARSINI, M. and SHAKED, M. (2005). Some notions of multivariate positive dependence. *Insurance Math. Econom.* **37** 13–26. MR2156593

[8] DAWID, A. P. (1979). Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **41** 1–31. MR0535541

[9] DAWID, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.* **8** 598–617. MR0568723

[10] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

[11] DJOLONGA, J. and KRAUSE, A. (2015). Scalable variational inference in log-supermodular models. In *Proceedings of the International Conference on Machine Learning (ICML)*. Available at ArXiv:1502.06531.

[12] EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer, New York. MR1880319

[13] ESARY, J. D., PROSCHAN, F. and WALKUP, D. W. (1967). Association of random variables, with applications. *Ann. Math. Stat.* **38** 1466–1474. MR0217826

[14] FORTUIN, C. M., KASTELEYN, P. W. and GINIBRE, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* **22** 89–103. MR0309498

[15] HØJSGAARD, S. and LAURITZEN, S. L. (2008). Graphical Gaussian models with edge and vertex symmetries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 1005–1027. MR2530327

[16] HOLLAND, P. W. and ROSENBAUM, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Statist.* **14** 1523–1543. MR0868316

[17] JOAG-DEV, K. (1983). Independence via uncorrelatedness under certain dependence structures. *Ann. Probab.* **11** 1037–1041. MR0714966

[18] JOE, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal.* **97** 2177–2189. MR2301633

[19] JONES, B. and WEST, M. (2005). Covariance decomposition in undirected Gaussian graphical models. *Biometrika* **92** 779–786. MR2234185

[20] KARLIN, S. and RINOTT, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10** 467–498. MR0599685

[21] KARLIN, S. and RINOTT, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra Appl.* **52/53** 419–438. MR0709364

[22] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Clarendon, Oxford. MR1419991

[23] LEBOWITZ, J. L. (1972). Bounds on the correlations and analyticity properties of ferromagnetic Ising spin systems. *Comm. Math. Phys.* **28** 313–321. MR0323271

[24] LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Stat.* **37** 1137–1153. MR0202228

[25] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

[26] MÜLLER, A. and SCARSINI, M. (2005). Archimedean copulae and positive dependence. *J. Multivariate Anal.* **93** 434–445. MR2162647

[27] NEWMAN, C. M. (1983). A general central limit theorem for FKG systems. *Comm. Math. Phys.* **91** 75–80. MR0719811

[28] NEWMAN, C. M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables. In *Inequalities in Statistics and Probability* (*Lincoln*, *Neb.*, 1982). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **5** 127–140. IMS, Hayward, CA. MR0789244

[29] OSTROWSKI, A. (1937). Über die determinanten mit überwiegender Hauptdiagonale. *Comment. Math. Helv.* **10** 69–96. MR1509568

[30] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA. MR0965765

[31] PEARL, J. and PAZ, A. (1987). Graphoids: A graph based logic for reasoning about relevancy relations. In *Advances in Artificial Intelligence*, *Vol. II* (B. D. Boulay, D. Hogg and L. Steel, eds.) 357–363. North-Holland, Amsterdam.

[32] PETERS, J. (2015). On the intersection property of conditional independence and its application to causal discovery. *J. Causal Inference* **3** 97–108.

[33] PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252. MR1611693

[34] REICHENBACH, H. (1956). *The Direction of Time*. Univ. California Press, Berkeley, CA.

[35] ROYSTON, P., ALTMAN, D. G. and SAUERBREI, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat. Med.* **25** 127–141. MR2222078

[36] SADEGHI, K. and LAURITZEN, S. (2014). Markov properties for mixed graphs. *Bernoulli* **20** 676–696. MR3178514

[37] SAN MARTÍN, E., MOUCHART, M. and ROLIN, J.-M. (2005). Ignorable common information, null sets and Basu's first theorem. *Sankhyā* **67** 674–698. MR2283009

[38] SARKAR, T. K. (1969). Some lower bounds of reliability. Tech. Report, No. 124, Dept. Operations Research and Dept. Statistics, Stanford Univ., Stanford, CA.

[39] SLAWSKI, M. and HEIN, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra Appl*. **473** 145–179. MR3338330

[40] STEEL, M. and FALLER, B. (2009). Markovian log-supermodularity, and its applications in phylogenetics. *Appl*. *Math*. *Lett*. **22** 1141–1144. MR2523016

[41] ŠTEMBERA, Z., ZNAMENÁČEK, K. and POLÁČEK, K. (2012). *High Risk Pregnancy and Child*. Springer, Berlin.

[42] STUDENÝ, M. (2005). *Probabilistic Conditional Independence Structures*. Springer, London. MR3183760

[43] WERMUTH, N. (2012). Traceable regressions. *Int*. *Stat*. *Rev*. **80** 415–438. MR3006154

[44] WERMUTH, N. (2015). Graphical Markov models, unifying results and their interpretation. In *Wiley Statsref*: *Statistics Reference Online*. Available at ArXiv:1505.02456.

[45] WERMUTH, N. and COX, D. R. (1998). On association models defined over independence graphs. *Bernoulli* **4** 477–495. MR1679794

[46] WERMUTH, N. and MARCHETTI, G. M. (2014). Star graphs induce tetrad correlations: For Gaussian as well as for binary variables. *Electron*. *J*. *Stat*. **8** 253–273. MR3189555

[47] WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester. MR1112133

[48] XIE, X., MA, Z. and GENG, Z. (2008). Some association measures and their collapsibility. *Statist*. *Sinica* **18** 1165–1183. MR2440406

[49] ZATOŃSKI, W., BECHER, H., LISSOWSKA, J. and WAHRENDORF, J. (1991). Tobacco, alcohol, and diet in the etiology of laryngeal cancer: A population-based case-control study. *Cancer Causes & Control* **2** 3–10.

[50] ZWIERNIK, P. (2016). *Semialgebraic Statistics and Latent Tree Models*. *Monographs on Statistics and Applied Probability* **146**. Chapman & Hall/CRC, Boca Raton, FL. MR3379921

S. FALLAT
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF REGINA
3737 WASCANA PKWY
REGINA, SASKATCHEWAN S4S 0A2
CANADA
E-MAIL: shaun.fallat@uregina.ca

S. LAURITZEN
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
2100 COPENHAGEN
DENMARK
E-MAIL: lauritzen@math.ku.dk

K. SADEGHI
STATISTICAL LABORATORY
CENTRE FOR MATHEMATICAL SCIENCES
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: k.sadeghi@statslab.cam.ac.uk

C. UHLER
IDSS / EECS DEPARTMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVE
CAMBRIDGE, MASSACHUSETTS 02139
USA
AND
INSTITUTE OF SCIENCE
AND TECHNOLOGY AUSTRIA
KLOSTERNEUBURG
AUSTRIA
E-MAIL: cuhler@mit.edu

N. Wermuth
Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg
Sweden
and
Johannes Gutenberg-University
Saarstrasse 21
55099 Mainz
Germany
E-mail: wermuth@chalmers.se

P. Zwiernik
Department of Economics and Business
and Barcelona GSE
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona
Spain
E-mail: piotr.zwiernik@upf.edu