TOTAL POSITIVITY IN MARKOV STRUCTURES

SHAUN FALLAT, STEFFEN LAURITZEN, KAYVAN SADEGHI, CAROLINE UHLER, NANNY WERMUTH, AND PIOTR ZWIERNIK

Abstract. 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 semigraphoid 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 and positive association. *Positive dependence* usually refers to two random variables that have a positive covariance, but other definitions of positive dependence have been proposed as well; see [23] for an overview. A random vector X is said to be *positively associated* if $\mathrm{cov}(f(X), g(X)) \geq 0$ for any two non-decreasing functions f and g [11]. The notion of positive association has important applications in probability theory and statistical physics; see, for example, [25, 26].

However, this notion of positive association is difficult to verify in any given application unless there is substantive knowledge already a-priori. The celebrated FKG theorem, formulated by Fortuin, Kasteleyn, and Ginibre in [12], introduces a simpler notion and states that X is positively associated if its 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

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

where $x \wedge y$ and $x \vee y$ denote the element-wise minimum and maximum.

Date: October 5, 2015.

2010 *Mathematics Subject Classification.* 60E15, 62H99, 15B48.

Key words and phrases. Concentration graph, conditional Gaussian distribution, faithfulness, graphical models, log-linear interactions, Markov property.

SF was supported in part by an NSERC Discovery Research Grant, KS 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), CU by the Austrian Science Fund (FWF) Y 903-N35, and PZ by the European Union Seventh Framework Programme PIOF-GA-2011-300975.

These inequalities are often easier to check in applications. Furthermore, all known definitions of positive dependence are implied by the MTP_2 constraints; see for example [6] 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 [21, 40].

MTP_2 distributions have also played an important role in the study of ferromagnetic Ising models, i.e., distributions of binary variables where all interaction potentials are pairwise and non-negative. It has been noted in [31] 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 [9] for an optimization viewpoint. The special case of Gaussian distributions was studied by Karlin and Rinott [20] and very recently by Slawski and Hein [35] from a machine learning perspective.

Consequences of MTP_2 distributions with respect to marginal and mutual independences were studied by Lebowitz [22] and Newman [26]. 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 3.6 and Theorem 5.4 below). This is remarkable because covariances and correlations are the weakest types of measures of dependence, see [43]; even though they identify independence in Gaussian distributions, this is often not the case in other types of distribution.

In this paper, we discuss implications of the MTP_2 constraints for conditional independence and vice versa. This paper can be seen as a continuation of work by Sarkar [34] and, in particular, by Karlin and Rinott [19, 20]. They noted that the family of MTP_2 distributions is invariant 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 random variables [19, p. 469] and for Gaussian variables [20, Theorem 3].

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 non-negative [19, Proposition 3.10]. This result holds true also for non-homogeneous 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 [41]. Similar statements apply to binary latent class models [2, 14, 41] and to latent tree models, both in the Gaussian and in the binary setting [36, 45]. Furthermore, many data sets are in fact 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 discuss several instances where the MTP_2 property appears in practice. Section 5 analyzes MTP_2 distributions with respect to conditional independence relations. One of the main results in this paper is Theorem 5.2, which shows that any independence model generated by an MTP_2 distribution is a singleton-transitive compositional semigraphoid 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 5.4 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 $v \in \Gamma$.

All distributions are assumed to have densities with respect to the product measure $\mu = \otimes_{v\in V}\mu_v$, 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*.

Finally, we introduce some definitions related to graphs: An *undirected graph* $G = (V, E)$ consists of a 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 $(v_0, v_1, \ldots, v_k)$ such that $v_iv_{i+1} \in E$ for all $i = 0, \ldots, k-1$ and no node is repeated, i.e., $v_i \neq v_j$ for all $i, j \in \{0, 1, \ldots, k\}$ 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 nodes $u, v \in V$ are *connected* if there is a path between u and v; a graph is *connected* if all pairs of nodes are connected. A graph is *complete* if all possible edges are present. In addition, two nodes $v, w \in V$ are *separated* by $S \subset V \setminus \{v, w\}$ if every path between v and w passes through a node in S. Finally, 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.

3. Basic properties and positive dependence

We start this section by formally introducing MTP_2 distributions, and we then discuss various basic properties of such distributions. We define the coordinate-wise minimum and maximum as

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

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

$$f(x)f(y) \quad \leq \quad f(x \wedge y)f(x \vee y) \qquad \text{for all } x, y \in \mathcal{X}. \tag{1}$$

For $|V| = 2$, a function that is MTP_2 is simply called *totally positive of order two* (TP_2) [19]. 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 defined on an open interval on the real line.

A basic property of MTP_2 distributions is that it is preserved under increasing transformations. We start 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 $\phi_v : \mathcal{X}_v \to \mathbb{R}$ is strictly increasing. If the distribution of X on $\mathcal{X}$ is* MTP_2*, then the distribution of $Y = \phi(X)$ is* MTP_2*.*

Proof. Without loss of generality we assume that $\mathcal{X}$ is either discrete or continuous. If $\mathcal{X}$ is discrete, the statement of the theorem is straightforward. If $\mathcal{X}$ is continuous, we use the following fact from [19, Equation (1.13)]: Let $a_v : \mathbb{R} \to \mathbb{R}$ be positive and let $b_v : \mathbb{R} \to \mathbb{R}$ be non-decreasing. If $f : \mathbb{R}^V \to \mathbb{R}$ is MTP_2, then the function

$$g(y) = f\{b_v(y_v), v \in V\} \prod_{v \in V} a_v(y_v), \tag{2}$$

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 (2) that Y is MTP_2. □

A similar result holds more generally for non-decreasing transformations. We here refrain from proving this in full generality; in the following proposition we prove that the MTP_2 property is preserved under non-decreasing transformations in the discrete setting.

Proposition 3.2. *Let X be a random vector taking values in a discrete space $\mathcal{X}$. Let $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathcal{X}_v \to \mathbb{R}$ is non-decreasing. If the distribution of X on $\mathcal{X}$ is* MTP_2*, then the distribution of $Y = \phi(X)$ is* MTP_2*.*

Proof. We have $\mathbb{P}(Y = y) = \int_{\phi^{-1}(y)} d\mu$. Note that

$$\phi^{-1}(y_1) \wedge \phi^{-1}(y_2) = \phi^{-1}(y_1 \wedge y_2), \quad \phi^{-1}(y_1) \vee \phi^{-1}(y_2) = \phi^{-1}(y_1 \vee y_2),$$

where for two sets A, B,

$$A \wedge B = \{a \wedge b \mid a \in A, b \in B\}, \quad A \vee B = \{a \vee b \mid a \in A, b \in B\}.$$

Hence, we can apply [19, Corollary 2.1] to obtain that Y is MTP_2. □

As we prove in the following result, the MTP_2 property is also 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 $\mathcal{X}_i = \{i_1, \ldots, i_p\}$ then $\mathcal{X}'_i = \{\{i_1, \ldots, i_j\}, i_{j+1}, \ldots, i_k, \{i_{k+1}, \ldots, i_p\}\}$ is a monotone coarsening.

Proposition 3.3. *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 [19, Proposition 3.2]. Property (iii) is an instance of a non-decreasing transformation and follows from Proposition 3.2. □

As we will see in Section 5, Section 6 and Section 7, the 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 [32]. This may still be true for MTP_2 distributions, see Example 6.2 below, but the coarsening property (iii) implies that associations cannot be distorted to 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

$$f(x)f(y) \neq 0 \qquad \text{implies} \qquad f(z) \neq 0 \text{ for any } x \wedge y \leq z \leq x \vee y. \tag{3}$$

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 [19, Proposition 2.1] prove the following result.

Proposition 3.4. *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 *coordinate-wise connected* [30], 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 3.5. Consider a binary $2 \times 2 \times 2$ distribution, where the support only misses the entries $(1,0,0)$ and $(1,0,1)$. All MTP_2 distributions must satisfy the nine inequalities in (6) below. In this example there are only two TP_2 constraints, namely

$$p_{000}p_{011} \geq p_{010}p_{001} \quad \text{and} \quad p_{010}p_{111} \geq p_{110}p_{011}. \tag{4}$$

These two inequalities on probabilities translate into conditions on conditional odds-ratios for $(1,3)$ and $(2,3)$: they imply positive, non-vanishing dependences for $(1,3)$ given 2 and for $(2,3)$ given 1, with the two odds-ratios being larger than one. Multiplying these the two inequalities in (4) yields

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

which after dividing both sides by $p_{010}p_{011}$ gives the only non-trivial MTP_2 constraint

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

Hence, in this case pairwise TP_2 constraints imply the MTP_2 property even though the distribution does not have interval support. □

As mentioned already in Section 1, if X is MTP_2, then X is *positively associated*, i.e.,

$$\operatorname{cov}\{f(X), g(X)\} \geq 0 \tag{5}$$

for any non-decreasing functions f and g. For discrete distributions this follows by the FKG theorem [12], or, more generally, by the four functions theorem by Ahlswede and Daykin [1]. The general case was proved by Sarkar [34]. The following result, first proven by Lebowitz [22], shows that the independence structure for positively associated vectors is encoded already in the covariance matrix; see also [17, 26].

Theorem 3.6 (Corollary 3, [26])**.** *Let X be positively associated. Then X_A is independent of X_B if and only if* $\operatorname{cov}(X_i, X_j) = 0$ *for all $i \in A$ and $j \in B$.*

In Section 5, we study conditional independence models for MTP_2 distributions. Interestingly, we will show in Theorem 5.4 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 dependent. This means in particular that for MTP_2 distributions, all marginal independences also hold when conditioning on any variable; the general version of this property will be termed *upward-stability*, see Section 5.

4. Examples of MTP_2 distributions in theory and practice

In this section we focus on multivariate Gaussian and binary MTP_2 distributions. Although the MTP_2 property may appear very restrictive, we want to suggest that MTP_2 distributions are important in practice and in fact appear in real data sets. We illustrate this with several examples.

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 [20], i.e.,

(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 [27] 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 showed that if X is MTP_2, then it is positively associated. Therefore, for MTP_2 Gaussian distributions we have 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 5.4 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 $1 \perp\!\!\!\perp 2 \mid 3$, then by a symmetry argument precisely 25% of such distributions are MTP_2. If, in addition, we impose $1 \perp\!\!\!\perp 3 \mid 2$ — 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 now discuss a prominent data set consisting of the examination marks of 88 students in five different mathematical subjects. The data were reported in [24] and analyzed, for example, in [10, 13, 42]. The inverse of the sample covariance matrix is displayed in Table 1. This matrix is very close to being an M-matrix with only one non-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

TABLE 1. Empirical partial correlations (below the diagonal) and concentrations ($\times 1000$, 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

$|V| = 3$, binary MTP_2 distributions must satisfy the following 9 inequalities

$$\begin{array}{lll} p_{011}p_{000} \geq p_{010}p_{001} & p_{101}p_{000} \geq p_{100}p_{001} & p_{110}p_{000} \geq p_{100}p_{010} \\ p_{111}p_{100} \geq p_{110}p_{101} & p_{111}p_{010} \geq p_{110}p_{011} & p_{111}p_{001} \geq p_{101}p_{011} \\ p_{111}p_{000} \geq p_{100}p_{011} & p_{111}p_{000} \geq p_{010}p_{101} & p_{111}p_{000} \geq p_{001}p_{110}. \end{array} \tag{6}$$

The first two rows correspond to the inequalities $p_{x\wedge y}p_{x\vee y} \geq p_x p_y$ as in (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 non-negative. By Proposition 3.4, 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

$$p_{111}p_{000} - p_{010}p_{101} = \frac{p_{000}}{p_{001}}(p_{111}p_{001} - p_{101}p_{011}) + \frac{p_{101}}{p_{001}}(p_{011}p_{000} - p_{010}p_{001}).$$

In order to verify whether a general positive distribution is MTP_2, one needs to check $\binom{|V|}{2} \cdot 2^{|V|-2}$ 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 which gives the conditions for log-linear parameters of general discrete distributions.

Similarly as for Gaussian distributions, the MTP_2 hypothesis is also very restrictive in the binary setting when no further conditional independences 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:

$\lvert V\rvert$	3	4
MTP_2	2195	0

Like 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 $1 \perp\!\!\!\perp 2 \mid 3$ are MTP_2. If, in addition, we have $1 \perp\!\!\!\perp 3 \mid 2$ 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 [41, 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 start with 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 [41, 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 [38, p. 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 (6). 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 on five binary random variables. This is a subset of data from a Polish case-control study on laryngeal cancer [44]. 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 $0, 1$ 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).

For the variables in this order, the level combinations, the corresponding observed and fitted counts under a well-fitting model are, respectively,

$$\begin{bmatrix} 00000 & 10000 & 01000 & 11000 \\ 00100 & 10100 & 01100 & 11100 \\ 00010 & 10010 & 01010 & 11010 \\ 00110 & 10110 & 01110 & 11110 \\ 00001 & 10001 & 01001 & 11001 \\ 00101 & 10101 & 01101 & 11101 \\ 00011 & 10011 & 01011 & 11011 \\ 00111 & 10111 & 01111 & 11111 \end{bmatrix} : \begin{bmatrix} 85 & 11 & 5 & 6 \\ 10 & 1 & 1 & 2 \\ 46 & 15 & 3 & 7 \\ 7 & 2 & 2 & 5 \\ 51 & 27 & 7 & 18 \\ 4 & 6 & 1 & 4 \\ 73 & 36 & 5 & 30 \\ 5 & 9 & 3 & 6 \end{bmatrix}, \begin{bmatrix} 85.88 & 9.87 & 7.30 & 6.35 \\ 9.27 & 1.70 & 1.55 & 2.08 \\ 47.59 & 14.31 & 4.19 & 9.21 \\ 5.32 & 2.47 & 0.89 & 3.02 \\ 51.70 & 27.13 & 4.55 & 17.46 \\ 5.78 & 4.68 & 0.97 & 5.73 \\ 70.57 & 39.96 & 6.22 & 25.72 \\ 7.89 & 6.89 & 1.32 & 8.43 \end{bmatrix}.$$

This well-fitting model yields an overall likelihood-ratio chi-square of 13.6 with 19 degrees of freedom, a concentration graph with cliques: $\{1,2,3\}$ and $\{1,4,5\}$, and the marginal tables in the set $\{\{1,2\},\{1,3\},\{2,3\},\{1,4,5\}\}$ as minimal sufficient statistics. 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, $\text{or}(I,J)$, and fitted conditional odds-ratios given the remaining variables, $\text{or}(I,J\,|\,R)$:

variable pair	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
observed $\text{or}(I,J)$	7.6	1.9	1.7	3.0	2.3	1.4	2.0	1.3	0.9	2.0
fitted $\text{or}(I,J)$						1.3	1.6	1.1	1.2	
fitted $\text{or}(I,J\,\|\,R)$	7.3	1.5	2.5	4.4	1.9	1	1	1	1	*)

*) 2.4 for controls and 1.02 for cases

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, $\text{or}(I,J\,|\,R)$, are less than 1. However in the well-fitting model we have $\text{or}(I,J\,|\,R) \geq 1$ for all 80 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 dependences for all variable pairs.

Given the two cliques of the graph, one sees that prediction of the 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 subjects only shows for controls but not for the cases. In combination with the fact that $\text{or}(1,5\,|\,R) = 4.4$, 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 [37].

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*).

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 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 [7, 21], i.e. 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

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

Probabilistic independence models are always semi-graphoids [29]. If, for example, P has a strictly positive density f, the induced independence model is always a graphoid; see e.g. Proposition 3.1 in [21]. More generally, if f is continuous, Peters [30] showed that the induced independence model is a graphoid if and only if the support is *coordinate-wise connected*, i.e., 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.

Examples of discrete distributions violating one of (S5), (S6), or (S7) have been given in [39]. We will prove in this section that, under weak assumptions, also independence models generated by MTP_2 distributions satisfy all the above properties (S1)-(S8).

We now prove that any induced independence model of an MTP_2 distribution is an upward-stable and singleton-transitive compositional semigraphoid. First, note that by Proposition 3.3 the MTP_2 property is closed under marginalization and conditioning. Therefore,

$$\operatorname{cov}\{\phi(X_u),\psi(X_v)\,|\,X_C\}\geq 0 \qquad \text{a.s.} \tag{7}$$

for all $u,v\in V$ with $u\neq v$, $C\subseteq V\setminus\{u,v\}$, and any non-decreasing functions $\phi,\psi:\mathbb{R}\to\mathbb{R}$. The following related result was first proved in [34, Section 3.1] (see also [19, Theorem 4.1]).

Proposition 5.1. *Let X be MTP_2. Then for any subset $A\subseteq V$ and any non-decreasing function $\varphi:\mathcal{X}_A\to\mathbb{R}$, the conditional expectation*

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

is non-decreasing in $x_{V\setminus A}$.

We now state and prove the main result in this section.

Theorem 5.2. *Any independence model $\mathcal{J}(P)$ induced by an MTP_2 distribution P is an upward-stable and singleton-transitive compositional semigraphoid.*

Proof. First, note that any probabilistic independence model is a semi-graphoid [28]. 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 $1\perp\!\!\!\perp 2\,|\,3$.

Independence of X_1 and X_2 implies that

$$\begin{aligned} 0 &= \operatorname{cov}(X_1,X_2)\\ &= \operatorname{cov}(\mathbb{E}(X_1\,|\,X_3),\mathbb{E}(X_2\,|\,X_3))+\mathbb{E}(\operatorname{cov}(X_1,X_2\,|\,X_3)). \end{aligned}$$

By Proposition 5.1, $\mathbb{E}(X_1\,|\,X_3=x_3)$ and $\mathbb{E}(X_2\,|\,X_3=x_3)$ are non-decreasing functions of x_3 and hence their covariance is non-negative by (5). Moreover, it follows from (7) that $\operatorname{cov}(X_1,X_2\,|\,X_3=x_3)\geq 0$ for almost all x_3, and thus its expectation is non-negative. This means that we expressed zero as a sum of two non-negative terms and thus both terms must be zero. This implies that $\operatorname{cov}(X_1,X_2\,|\,X_3=x_3)=0$ for almost all x_3. Hence, by Theorem 3.6, we obtain that $1\perp\!\!\!\perp 2\,|\,3$.

Next we prove that $\mathcal{J}(P)$ has the composition property. By the previous argument, $A \perp\!\!\!\perp B \,|\, C$ implies $A \perp\!\!\!\perp B \,|\, C \cup D$. Hence, composition follows directly from contraction.

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 $1 \perp\!\!\!\perp 2$. Then $1 \perp\!\!\!\perp 2 \,|\, 3$ by upward-stability and hence

$$\begin{aligned} 0 &= \operatorname{cov}(X_1, X_2) \\ &= \operatorname{cov}(\mathbb{E}(X_1 \,|\, X_3), \mathbb{E}(X_2 \,|\, X_3)) + \mathbb{E}(\operatorname{cov}(X_1, X_2 \,|\, X_3)) \\ &= \operatorname{cov}(\mathbb{E}(X_1 \,|\, X_3), \mathbb{E}(X_2 \,|\, X_3)). \end{aligned}$$

Then by Theorem 5.1, both $\mathbb{E}(X_1 \,|\, X_3)$ and $\mathbb{E}(X_2 \,|\, X_3))$ are non-decreasing functions. Since their covariance is 0, $\mathbb{E}(X_1 \,|\, X_3)$ or $\mathbb{E}(X_2 \,|\, X_3)$ is constant almost surely. Assuming that $p_3(x_3) > 0$ for almost all x_3 (otherwise we collapse the state space) implies that $\operatorname{cov}(X_1, X_3) = 0$ or $\operatorname{cov}(X_2, X_3) = 0$. Then by Theorem 3.6 we obtain that $1 \perp\!\!\!\perp 3$ or $2 \perp\!\!\!\perp 3$. □

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 5.3. 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 [30], any MTP_2 distribution with continuous density and coordinate-wise connected support is an upward-stable singleton-transitive compositional graphoid.

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

Theorem 5.4. *Let the distribution of X be MTP_2. Then X can be decomposed into independent components such that within each component all variables are mutually marginally dependent.*

Proof. By Theorem 3.6 it suffices to prove that the covariance matrix of X is block diagonal with strictly positive entries in each block. We write $u \sim v$ if the covariance between X_u and X_v is non-zero 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 $u \not\sim w$, we have $\sigma_{uw} = 0$ and thus $u \perp\!\!\!\perp w$. Using upward-stability from Theorem 5.2 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 $v \sim w$. □

6. Faithfulness and total positivity

In the following, we 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

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

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 *concentration graph* of P is the undirected graph $G(P) = (V, E(P))$ with

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

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

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

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

Note that for MTP_2 distributions, Proposition 3.3 and Theorem 3.6 imply that the concentration graph $G(P)$ for an MTP_2 distribution is equal to the *partial correlation graph* $\Pi(P)$, where

$$uv \notin \Pi(P) \iff \rho_{uv \,|\, V\setminus\{u,v\}} = 0.$$

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 coordinate-wise 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 [33]. We prove in the following result that if the pairwise-global equivalence is already established, then P is in fact faithful to $G(P)$ and thus also to $\Pi(P)$.

Theorem 6.1. *If P is MTP_2 and has coordinate-wise connected support, then P is faithful to its concentration and partial correlation graph $G(P) = \Pi(P)$.*

Proof. From Theorem 5.2 we get that the independence model $\mathcal{J}(\mathcal{P})$ is a graphoid. Thus by Theorem 3.7 in [21] it follows that P is globally Markov w.r.t. $G(P)$.

Next we prove the converse: 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 $u \not\perp\!\!\!\perp v \mid V \setminus \{u, v\}$ and hence by upward-stability as shown in Theorem 5.2, $u \not\perp\!\!\!\perp v \mid C$ 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 $u = v_1, v_2, \ldots, v_r = v$ such that $v_k \notin C$ for all $k = 1, \ldots, r$, and $v_k v_{k+1} \in E$ for all $k = 1, \ldots, r-1$. By the previous argument we obtain that $v_k \not\perp\!\!\!\perp v_{k+1} \mid C$ for all $k = 1, \ldots, r-1$. By singleton-transitivity, $v_1 \not\perp\!\!\!\perp v_2 \mid C$ and $v_2 \not\perp\!\!\!\perp v_3 \mid C$ imply that $v_1 \not\perp\!\!\!\perp v_3 \mid C$. Repeating this argument yields $u \not\perp\!\!\!\perp v \mid C$ and hence $A \not\perp\!\!\!\perp B \mid C$. □

Notice that if X has coordinate-wise connected support, then Theorem 5.4 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 5.4 obviously holds. However, the theorem is still interesting as it covers cases that Theorem 6.1 does not cover; such as that of Example 5.3.

In Section 3 we postponed to show that the stability of the MTP_2 property under coarsening as established in (iii) of Proposition 3.3 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 $I \perp\!\!\!\perp K \mid J$ 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

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

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 non-negativity 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.

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

Proof. By the characterization of MTP_2 Gaussian distributions given in Section 4.1 it suffices to prove that it is possible to construct a positive definite symmetric M-matrix with any given pattern of zeros.

Let $G = (V, E)$ be an undirected graph with $|V| = m$, and let A be the $(0, 1)$-adjacency matrix associated with G, i.e., $A = [a_{uv}]$ with

$$a_{uv} = \begin{cases} 1 & \text{if } uv \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let Δ denote the maximum degree in G. Define the matrix $B = \Delta I - A$. Then B is a real symmetric matrix, and B is an M-matrix (see also [16, Section 2.5]). To see the latter claim, it suffices to verify that the eigenvalues of B are all non-negative. For this, we will make use of Gershgorin's circle theorem (see [15, Chapter 6]). Recall that the eigenvalues of B are all contained in the union of discs as follows:

$$\bigcup_{v \in V} \left\{ |z - b_{vv}| \leq \sum_{u \neq v} |b_{uv}| \right\}.$$

By construction of B, we know that $\Delta = b_{vv} \geq \sum_{u \neq v} |b_{uv}| = \sum_{u \neq v} a_{uv}$, since the latter sum is simply the degree of vertex v. Hence, since $b_{vv} \geq 0$, all of these discs lie in the right-half plane. In other words, B is a real symmetric positive semidefinite matrix. Since all of the off-diagonal entries are non-positive, it follows that B is an M-matrix. Finally, since A respects the graph G, so does B. Now for any $\epsilon > 0$, $K = \epsilon I + B$ is a positive definite M-matrix which respects the graph G. □

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 random variable 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 [8]). 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$_2$. For notational simplicity let $A \cup B = V$ and $A \cap B = \{v\}$. Then

$$(f \star g)(x \wedge y) \cdot (f \star g)(x \vee y) \quad = \quad \frac{f(x \wedge y)g(x \wedge y)}{h(x_v \wedge y_v)} \frac{f(x \vee y)g(x \vee y)}{h(x_v \vee y_v)},$$

and since $\{x_v, y_v\} = \{x_v \wedge y_v, x_v \vee y_v\}$, we obtain

$$(f \star g)(x \wedge y) \cdot (f \star g)(x \vee y) \quad = \quad \frac{f(x \wedge y)g(x \wedge y)f(x \vee y)g(x \vee y)}{h(x_v)h(y_v)}.$$

Now we use the fact that f and g are MTP$_2$ to obtain

$$\begin{aligned}(f \star g)(x \wedge y) \cdot (f \star g)(x \vee y) \quad &\geq \quad \frac{f(x)g(x)f(y)g(y)}{h(x_v)h(y_v)} \\ &= \quad (f \star g)(x) \cdot (f \star g)(y),\end{aligned}$$

which completes the proof. □

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 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 (6) 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:

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

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

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

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

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

P is not MTP_2. □

As a direct consequence of Proposition 7.1 we obtain the following result for *decomposable graphs*, which are equivalent to graphs where there is no cycle of length more than 3 such that all its non-neighboring nodes (on the cycle) are not adjacent (see e.g. [21] 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 non-decomposable 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

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

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 non-negative. In this example,

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

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 non-decomposable graphs, namely by using MTP_2 functions over the edges instead of MTP_2 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 3.4 it suffices to show the MTP_2 condition 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

$$\psi_{uv}((x \wedge y)_{uv})\, \psi_{uv}((x \vee y)_{uv}) \prod_{st \in E_u \cup E_v} \psi_{st}((x \wedge y)_{st})\, \psi_{st}((x \vee y)_{st}) \geq$$
$$\psi_{uv}(x_{uv})\, \psi_{uv}(y_{uv}) \prod_{st \in E_u \cup E_v} \psi_{st}(x_{st})\, \psi_{st}(y_{st}).$$

All other terms cancel out because of the assumption that $x_w = y_w$ for all $w \in V \setminus \{u, v\}$. 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

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

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

Next consider the case where $uv \notin E$. By the same argument one finds 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 $X = (X_v)_{v \in V}$ be a random vector with values in $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, where $\mathcal{X}_v = \{0, \ldots, d_v\}$. Let $\mathcal{A}$ be a *generating class* of pairwise incomparable subsets of V and $\mathcal{D}$ the associated *simplicial complex*, i.e. $D \in \mathcal{D}$ if and only if there is an $A \in \mathcal{A}$ so that $D \subseteq A$. The hierarchical log-linear model generated by $\mathcal{D}$ is the set of positive probabilities $p = \{p(x)\}_{x \in \mathcal{X}}$ on $\mathcal{X}$ such that

$$\log p(x) \;=\; \sum_{D \in \mathcal{D}} \theta_D(x_D). \tag{8}$$

Here $\theta_D(x_D)$ are the *log-linear interactions* among variables in D, that is functions on $\mathcal{X}$ such that $\theta_D(x) = \theta_D(x_D)$. To assure that the representation is unique we also require that $\theta_D(x_D) = 0$ whenever $x_d = 0$ for some $d \in D$. With this convention, the sum in (8) 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 shorter notation, namely $\theta_D(\mathbf{1}_D) := \theta_D$ for all $D \in \mathcal{D}$.

For a fixed pair $u, w \in V$, define a function γ_{uw} on $\mathcal{X}$ by

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

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

A function g on $\mathcal{X}$ is *supermodular* if

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

Thus a function g is supermodular if and only if $\exp(g)$ is MTP_2.

Denote by $\mathcal{X}^A$ the set of all $x \in \mathcal{X}$ with $S(x) = A$ Then we obtain the following result.

Theorem 7.6. *Let P be a distribution of X. Then P is MTP_2 if and only if for any given $u, w \in V$ the function γ_{uw} is non-negative, and, over each $\mathcal{X}^A$ for $|A| \geq 2$, it is non-decreasing and supermodular.*

Proof. By Proposition 3.4, to check if p is MTP_2, it suffices to check if

$$\log p(x \wedge y) + \log p(x \vee y) - \log p(x) - \log p(y) \geq 0$$

for all $x, y \in \mathcal{X}$ that differ only in two entries. Suppose that there exists $u, w \in V$ 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 above is trivially satisfied. By (8), the expression $\log p(x \wedge y) + \log p(x \vee y) - \log p(x) - \log p(y) \geq 0$ becomes

$$\sum_{D \in \mathcal{D}} \big(\theta_D((x \wedge y)_D) + \theta_D((x \vee y)_D) - \theta_D(x_D) - \theta_D(y_D)\big) \quad \geq \quad 0, \tag{9}$$

where $\mathcal{D}$ is the set of all non-empty subsets of V.

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 (9) are zero. It follows that (9) can be rewritten as

$$\sum_{\{u,w\} \subseteq D \subseteq A} \big(\theta_D((x \wedge y)_D) + \theta_D((x \vee y)_D) - \theta_D(x_D) - \theta_D(y_D)\big) \quad \geq \quad 0. \tag{10}$$

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 (10) 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) $S(x) = A \setminus \{u\}$, $S(y) = A \setminus \{v\}$, (b) either $S(x) = A \setminus \{u\}$, $S(y) = A$ or $S(x) = A$, $S(y) = A \setminus \{v\}$; and (c) $S(x) = S(y) = A$. In other words: (a) $x_u = y_w = 0$, (b) either $x_u = 0$, $y_w > 0$ or $x_u > 0$, $y_w = 0$; and (c) $x_u, y_w > 0$. In case (a) we have $\theta_D((x \wedge y)_D) = \theta_D(x_D) = \theta_D(y_D) = 0$ for every D containing $\{u, w\}$ and so

(10) becomes $\sum_{\{u,w\}\subseteq D\subseteq A}\theta_D((x\vee y)_D)\geq 0$. By choosing different pairs x, y, this can be equivalently rewritten as

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

and the sum on the left is precisely $\gamma_{uw}(x)$. In case (b), if $S(x)=A\setminus\{u\}$, $S(y)=A$, (10) becomes $\sum_{\{u,w\}\subseteq D\subseteq A}(\theta_D((x\vee y)_D)-\theta_D(y_D))\geq 0$, 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 description of binary MTP_2 distributions in [4].

Corollary 7.7. *Let P be a binary distribution with $\log p(x)=\sum_{D\in\mathcal{D}}\theta_D$ and the convention that $\theta_D=\theta_D(\mathbf{1}_D)$. Then P is* MTP_2 *if and only if*

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

Proof. In the binary case each $\mathcal{X}_A$ has only one element and so the only constraint from Theorem 7.6 is the non-negativity constraint. □

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 Gaussian $\mathcal{N}_\Gamma(\xi(i),\Sigma(i))$, 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

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

(see [21]). 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

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

For a CG-distribution we have the following conditions for the MTP_2 property.

Proposition 7.8. *A CG-distribution with canonical characteristics (g,h,K) is* MTP_2 *if and only if*

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

Proof. By Proposition 3.4, it suffices to check that

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

only in the case when (x,i) and (y,j) differ on two coordinates. Suppose first that $i=j$ and x,y differ on two coordinates. Then we equivalently need to show that

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

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

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

$$\log f(y, i) + \log f(x, j) \geq \log f(x, i) + \log f(y, j). \tag{11}$$

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

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

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

$$\nabla_y \log f(y, j) \geq \nabla_y \log f(y, i) \quad \text{for all } y \in \mathbb{R}^\Gamma, i < j \in \Delta,$$

which is equivalent to

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

Since on the left-hand side there is a linear function, this can hold for every y only if $K(j) = K(i)$ for every i, j and $h(i)$ is non-decreasing in i. It turns out that these conditions are also sufficient for (11) to hold.

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

$$(g(i \wedge j) + g(i \vee j) - g(i) - g(j)) + (h(i \wedge j) + h(i \vee j) - h(i) - h(j))^T x \geq 0. \tag{12}$$

This can hold for every x only if

$$h(i \wedge j) + h(i \vee j) - h(i) - h(j) = 0 \text{ for all } i, j. \tag{13}$$

Now if (13) holds, (12) holds if and only if $g(i)$ is super-modular. Finally, we make a log-linear expansion of h as in Section 6.2.1 of [21], namely

$$h(i)_v = \sum_{D \in \mathcal{D}} \eta_D(i)_v,$$

. Then Theorem 7.6 applies to the expansion of h. Since h satisfies (13), both of h and $-h$ are supermodular, and all γ^h_{uw} functions must be zero. Hence $h(i)$ is additive. □

Proposition 7.8 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 7.9. *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 non-decreasing;*
(iii) $\Sigma(i) = \Sigma$ *for all* i *and all elements of* Σ *are non-negative.*

Proof. If the CG distribution is MTP_2, (iii) follows directly from Proposition 7.8. The condition (i) follows since marginals of MTP_2 distributions are MTP_2 and (ii) follows from (ii) of Proposition 7.8 since $\xi(i) = \Sigma h(i)$ and Σ has only non-negative elements. □

Thus, MTP_2 CG-distributions are in particular *homogeneous* — $\Sigma(i)$ constant in i — and *mean-additive* [10, 21]. Note that the converse of Proposition 7.9 is not true since $\xi(i)$ can be non-increasing and $h = K\xi(i)$ decreasing, even when K is an M-matrix.

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

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

We then have the following alternative formulation of Proposition 7.8:

Corollary 7.10. *A CG-distribution is* MTP_2 *if and only if*

(i) *The functions* $\gamma^g_{uw}(i)$ *are supermodular and non-decreasing over each* $\mathcal{I}^A$*;*
(ii) $h(i)$ *is additive*
$$h(i) = \sum_{\delta\in\Delta} \alpha_\delta(i_\delta)$$
with non-decreasing components $\alpha_\delta(i_\delta)_v$*;*
(iii) *there exists an* M*-matrix* K *such that* $K(i) = K$ *for all* i*.*

Proof. Theorem 7.6 applies 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

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

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 coordinate-wise 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 methods for hypothesis testing and estimation under the MTP_2 constraint. 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.

Acknowledgements

This article represents research initiated at the American Institute of Mathematics workshop “Positivity, Graphical Models, and the Modeling of Complex Multivariate Dependencies”. 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.

References

[1] R. Ahlswede and D. E. Daykin, *An inequality for the weights of two families of sets, their unions and intersections*, Probability Theory and Related Fields, 43 (1978), pp. 183–185.

[2] E. S. Allman, J. A. Rhodes, B. Sturmfels, and P. Zwiernik, *Tensors of nonnegative rank two*, Linear Algebra Appl., 473 (2015), pp. 37–53.

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

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

[5] E. Bølviken, *Probability inequalities for the multivariate normal with non-negative partial correlations*, Scand. J. Statist., (1982), pp. 49–58.

[6] A. Colangelo, M. Scarsini, and M. Shaked, *Some notions of multivariate positive dependence*, Insurance Math. Econom., 37 (2005), pp. 13–26.

[7] A. P. Dawid, *Conditional independence in statistical theory (with discussion)*, J. Roy. Stat. Soc. Ser. B, 41 (1979), pp. 1–31.

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

[9] J. Djolonga and A. Krause, *Scalable variational inference in log-supermodular models*, arXiv preprint arXiv:1502.06531, (2015).
[10] D. Edwards, *Introduction to Graphical Modelling*, Springer Texts in Statistics, Springer-Verlag, New York, second ed., 2000.
[11] J. D. Esary, F. Proschan, and D. W. Walkup, *Association of random variables, with applications*, Ann. Math. Stat., 38 (1967), pp. 1466–1474.
[12] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre, *Correlation inequalities on some partially ordered sets*, Comm. Math. Phys., 22 (1971), pp. 89–103.
[13] S. Højsgaard and S. L. Lauritzen, *Graphical Gaussian models with edge and vertex symmetries*, J. R. Stat. Soc. Ser. B, 70 (2008), pp. 1005–1027.
[14] P. W. Holland and P. R. Rosenbaum, *Conditional association and unidimensionality in monotone latent variable models*, Ann. Statist., 14 (1986), pp. 1523–1543.
[15] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.
[16] ——, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1991.
[17] K. Joag-Dev, *Independence via uncorrelatedness under certain dependence structures*, Ann. Probab., 11 (1983), pp. 1037–1041.
[18] H. Joe, *Generating random correlation matrices based on partial correlations*, J. Multiv. Anal., 97 (2006), pp. 2177 – 2189.
[19] S. Karlin and Y. Rinott, *Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions*, J. Multiv. Anal., 10 (1980), pp. 467–498.
[20] ——, *M-matrices as covariance matrices of multinormal distributions*, Linear Algebra Appl., 52 (1983), pp. 419 – 438.
[21] S. L. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, United Kingdom, 1996.
[22] J. L. Lebowitz, *Bounds on the correlations and analyticity properties of ferromagnetic Ising spin systems*, Comm. Math. Phys., 28 (1972), pp. 313–321.
[23] E. L. Lehmann, *Some concepts of dependence*, Ann. Math. Statist., 37 (1966), pp. 1137–1153.
[24] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.
[25] C. M. Newman, *A general central limit theorem for FKG systems*, Comm. Math. Phys., 91 (1983), pp. 75–80.
[26] C. M. Newman, *Asymptotic independence and limit theorems for positively and negatively dependent random variables*, Lecture Notes-Monograph Series, 5 (1984), pp. 127–140.
[27] A. Ostrowski, *Über die Determinanten mit überwiegender Hauptdiagonale*, Commentarii Mathematici Helvetici, 10 (1937), pp. 69–96.
[28] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1988.
[29] J. Pearl and A. Paz, *Graphoids: A Graph Based Logic for Reasoning about Relevancy Relations*, in Advances in Artificial Intelligence, B. Boulay, D. Hogg, and L. Steel, eds., vol. II, North-Holland, Amsterdam, 1987, pp. 357–363.
[30] J. Peters, *On the intersection property of conditional independence and its application to causal discovery*, J. Causal Inference, 3 (2015), pp. 97–108.
[31] J. G. Propp and D. B. Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Random Structures Algorithms, 9 (1996), pp. 223–252.
[32] P. Royston, D. G. Altman, and W. Sauerbrei, *Dichotomizing continuous predictors in multiple regression: a bad idea*, Stat Med, 25 (2006), pp. 127–141.
[33] K. Sadeghi and S. L. Lauritzen, *Markov properties for mixed graphs*, Bernoulli, 20 (2014), pp. 676–696.
[34] T. K. Sarkar, *Some lower bounds of realiability*, Tech. Report, No. 124 (2nd ed.), Department of Operations Research and Department of Statistics, Stanford University, 1969.
[35] M. Slawski and M. Hein, *Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random field*, Linear Algebra Appl., 473 (2015), pp. 145–179.

[36] M. Steel and B. Faller, *Markovian log-supermodularity, and its applications in phylogenetics*, Appl. Math. Lett., 22 (2009), pp. 1141 – 1144.
[37] M. Studený, *Probabilistic Conditional Independence Structures*, Springer-Verlag, London, United Kingdom, 2005.
[38] Z. Štembera, K. Znamenáček, and K. Poláček, *High risk pregnancy and child*, Springer Science & Business Media, 2012.
[39] N. Wermuth, *Traceable regressions*, Int. Stat. Rev., 80 (2012), pp. 415–438.
[40] ——, *Graphical Markov models, unifying results and their interpretation.*, in Wiley Statsref: Statistics Reference Online, to appear; also on arXiv:1505.02456, John Wiley & Sons, 2014.
[41] N. Wermuth and G. M. Marchetti, *Star graphs induce tetrad correlations: for Gaussian as well as for binary variables*, Electron. J. Stat., 8 (2014), pp. 253–273.
[42] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester, 1990.
[43] X. Xie, Z. Ma, and Z. Geng, *Some association measures and their collapsibility*, Statist. Sinica, 18 (2008), pp. 1165–1183.
[44] W. Zatoński, H. Becher, J. Lissowska, and J. Wahrendorf, *Tobacco, alcohol, and diet in the etiology of laryngeal cancer: a population-based case-control study*, Cancer Causes & Control, 2 (1991), pp. 3–10.
[45] P. Zwiernik, *Semialgebraic Statistics and Latent Tree Models*, no. 146 in Monographs on Statistics and Applied Probability, Chapman & Hall, 2015.

Department of Mathematics and Statistics, University of Regina, Regina, SK, Canada

E-mail address: shaun.fallat@uregina.ca

Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

E-mail address: lauritzen@math.ku.dk

Statistical Laboratory, University of Cambridge, Cambridge, UK

E-mail address: k.sadeghi@statslab.cam.ac.uk

Massachusetts Institute of Technlogy, Cambridge, USA and Institute of Science and Technology Austria, Klosterneuburg, Austria

E-mail address: cuhler@mit.edu

Chalmers University of Technology, Gothenburg, Sweden and Johannes Gutenberg-University, Mainz, Germany

E-mail address: wermuth@chalmers.se

Department of Mathematics, University of Genoa, Genoa, Italy and Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain

E-mail address: piotr.zwiernik@gmail.com