Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

9 Multivariate Totally Positive Distributions

As we have seen so far, graphical models represent conditional indepedence relations in the joint probability distributions of random variables. We have also seen examples of structures in the distributions that are not represented by graphical models, e.g., jointly Gaussian distributions. Jointly Gaussian random variables have a nice property that they only have pairwise potentials, which is not represented by the structure of a graphical model, but allows for efficient inference. In this installment of the lecture notes, we will see another such example of a structure in the joint distribution that has nice implications for graphical models.

9.1 Multivariate totally positive (MTP_2) distributions

If many applications you observe random variables moving "together." Consider for example prices of stocks as random variables; you will often observe that the stock prices move "together." In many cases they go up together and go down together, although there is some slight differences in individual stocks. In this case, we say there is *positive dependence*. Totally positive distributions have been introduced to model positive dependence between random variables.

Definition 1. *A distribution* $p(\mathbf{x})$ *is* multivariate totally positive *of order 2 (MTP_2) if*

$$p(\mathbf{x})p(\mathbf{y}) \leq p(\mathbf{x} \wedge \mathbf{y})p(\mathbf{x} \vee \mathbf{y}), \tag{1}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}^N$.

Here, $\wedge$ means the *coordinate-wise* minimum between $\mathbf{x}$ and $\mathbf{y}$, and $\vee$ denotes the coordiante-wise maximum. More concretely,

$$\mathbf{x} \wedge \mathbf{y} = (x_1 \wedge y_1, x_2 \wedge y_2, \ldots, x_N \wedge y_N),$$
$$\mathbf{x} \vee \mathbf{y} = (x_1 \vee y_1, x_2 \vee y_2, \ldots, x_N \vee y_N).$$

For example, consider $\mathbf{x} = (1, 0)$ and $\mathbf{y} = (0, 1)$, then $\mathbf{x} \wedge \mathbf{y} = (0, 0)$, and $\mathbf{x} \vee \mathbf{y} = (1, 1)$.

Example 1. For a simple illustration, let us look at the 2-dimensional binary case $\mathbf{x}, \mathbf{y} \in \{0, 1\}^2$ and see what the constraint (1) means. Since (1) has to be satisfied for all $\mathbf{x}$ and $\mathbf{y}$, let us choose some example pairs of $\mathbf{x}$ and $\mathbf{y}$. Suppose we choose $\mathbf{x} = (0, 0)$ and $\mathbf{y} = (0, 1)$. This gives us

$$p(0, 0)p(0, 1) \leq p(0, 0)p(0, 1),$$

which is a trivial statement that is always satisfied. There are many trivial statements that we get for other $\mathbf{x}$'s and $\mathbf{y}$'s, but there is one nontrivial statement, namely, when $\mathbf{x} = (0, 1)$ and $\mathbf{y} = (1, 0)$:

$$p(0,1)p(1,0) \leq p(0,0)p(1,1).$$

So the probability of seeing $(0, 1)$ times the probability of seeing $(1, 0)$ must be less than or equal to that of $(0, 0)$ times $(1, 1)$. In words, the random variables are more likely to be in the same state than different states; they show positive dependence. In other words, the log-odds ratio is non-negative. This is the definition of MTP_2 distributions in the 2-dimensional binary setting.

For larger sample spaces, the definition (1) introduces a lot of constraints, so it is in general difficult to check (1) in a brute-force manner. For example, in the 3-dimensional binary case (i.e., $\mathbf{x} \in \{0,1\}^3$), (1) already introduces 9 inequalities:

$$\begin{aligned}
&p(0,0,1)p(1,1,0) \leq p(0,0,0)p(1,1,1), \quad p(0,1,0)p(1,0,1) \leq p(0,0,0)p(1,1,1),\\
&p(1,0,0)p(0,1,1) \leq p(0,0,0)p(1,1,1), \quad p(0,1,1)p(1,0,1) \leq p(0,0,1)p(1,1,1),\\
&p(0,1,1)p(1,1,0) \leq p(0,1,0)p(1,1,1), \quad p(1,0,1)p(1,1,0) \leq p(1,0,0)p(1,1,1),\\
&p(0,0,1)p(0,1,0) \leq p(0,0,0)p(0,1,1), \quad p(0,0,1)p(1,0,0) \leq p(0,0,0)p(1,0,1),\\
&p(0,1,0)p(1,0,0) \leq p(0,0,0)p(1,1,0).
\end{aligned} \tag{2}$$

Example 2. While MTP_2 is a very restrictive constraint, as seen by the many inequalities an MTP_2 distribution needs to satisfy already in the 3-dimensional binary setting in (2), there are datasets that satisfy MTP_2. One example of an empirical distribution that satisfies the set of inequalities in (2) is from a study on "Pregnancy and Child Development" performed by the German Research Foundation. They collected data points recording three symptoms that appear for pregnant women: edema (high body water retention), proteinuria (high amounts of urinary proteins), and hypertension (elevated blood pressure). The observed counts are:

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

and one can check that the counts satisfy (2); so the empirical distribution is MTP_2. This is quite remarkable, since were you to sample 3-dimensional binary distributions at random, only about 2% are MTP_2.

Although MTP_2 is difficult to check in general, this is in fact much easier to check than the condition people used to model positive dependence before MTP_2, namely *positive association.* A random vector $\mathbf{x}$ is positively associated if for any non-decreasing functions $\phi, \psi : \mathbb{R}^N \to \mathbb{R}$

$$\mathrm{cov}\{\phi(\mathbf{x}), \psi(\mathbf{x})\} \geq 0.$$

It can be shown [1,2] that MTP_2 implies positive association and this was in fact the reason why MTP_2 was introduced in the first place.

9.2 Gaussian MTP_2 distributions

Next, we discuss what it means for a Gaussian distribution to be MTP_2. It turns out that for Gaussians, we have a simple characterization of MTP_2 distributions, which comes from the following theorem:

Theorem 1 ([3,4])**.** *Let $p > 0$ be a strictly positive distribution with $f(x) = \log p(x)$. Then the following statements are equivalent:*

(a) p is MTP_2;

(b) p is MTP_2 for any pair of coordinates when the others are held fixed;

(c) $\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0$ for all $i \neq j$ and $\mathbf{x} \in \mathcal{X}^N$.

Proof. (a) $\implies$ (b): Since p is MTP_2, the definition (1) holds for any pair of coordinates when the others are fixed.

(b) $\iff$ (c): Assume without loss of generality that $i = 1$ and $j = 2$, and also let $\tilde{f}(x_1, x_2) = f(x_1, x_2, \mathbf{x}_{\mathcal{V}\setminus\{1,2\}})$ for any fixed value of $\mathbf{x}_{\mathcal{V}\setminus\{1,2\}}$. Then (b) is equivalent to

$$
\begin{aligned}
0 &\leq \log \frac{p(x_1 + h_1, x_2 + h_2, \mathbf{x}_{\mathcal{V}\setminus\{1,2\}}) p(x_1, x_2, \mathbf{x}_{\mathcal{V}\setminus\{1,2\}})}{p(x_1 + h_1, x_2, \mathbf{x}_{\mathcal{V}\setminus\{1,2\}}) p(x_1, x_2 + h_2, \mathbf{x}_{\mathcal{V}\setminus\{1,2\}})} \\
&\leq \tilde{f}(x_1 + h_1, x_2 + h_2) - \tilde{f}(x_1 + h_1, x_2) - \tilde{f}(x_1, x_2 + h_2) + \tilde{f}(x_1, x_2),
\end{aligned}
$$

for any $h_1, h_2 \geq 0$. Now, note that if we divide the right-hand side by $h_1 h_2$ and take the limit as $h_1, h_2 \to 0$, we get the second derivative $\frac{\partial^2 f}{\partial x_1 \partial x_2} \geq 0$.

(b) $\implies$ (a): We apply mathematical induction on the number of variables. Suppose that p is MTP_2 for any $(k-1)$-tuple of coordinates when the others are fixed. From this, we show that p is MTP_2 for any k-tuple of coordinates. Without loss of generality, denote

$$
\begin{aligned}
\mathbf{x} &= (x_1^*, \ldots, x_j^*, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N), \\
\mathbf{y} &= (y_1, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N), \\
\tilde{\mathbf{x}} &= (x_1^*, y_2 \ldots, y_j, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N), \\
\tilde{\mathbf{y}} &= (x_1^*, y_2, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N),
\end{aligned}
$$

where $x_i^* \geq y_i$ for $i = 1, \ldots, j$ and $x_i \leq y_i^*$ for $i = j+1, \ldots, k$. Then

$$
\begin{aligned}
&\frac{p(\mathbf{x} \vee \mathbf{y}) p(\mathbf{x} \wedge \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} \\
&= \frac{p(x_1^*, \ldots, x_j^*, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N) p(y_1, \ldots, y_j, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N)}{p(x_1^*, \ldots, x_j^*, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N) p(y_1, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N)} \\
&= \frac{p(x_1^*, \ldots, x_j^*, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N) p(x_1^*, y_2, \ldots, y_j, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N)}{p(x_1^*, \ldots, x_j^*, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N) p(x_1^*, y_2, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N)}
\end{aligned}
$$

$$\times \frac{p(x_1^*, y_2, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N) p(y_1, \ldots, y_j, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N)}{p(x_1^*, y_2, \ldots, y_j, x_{j+1}, \ldots, x_k, z_{k+1}, \ldots, z_N) p(y_1, \ldots, y_j, y_{j+1}^*, \ldots, y_k^*, z_{k+1}, \ldots, z_N)}$$

$$= \frac{p(\mathbf{x} \vee \tilde{\mathbf{y}}) p(\mathbf{x} \wedge \tilde{\mathbf{y}})}{p(\mathbf{x}) p(\tilde{\mathbf{y}})} \frac{p(\tilde{\mathbf{x}} \vee \mathbf{y}) p(\tilde{\mathbf{x}} \wedge \mathbf{y})}{p(\tilde{\mathbf{x}}) p(\mathbf{y})}$$

Note that the right-hand side of the equality has two multiplicative terms. There are only $k-1$ variables that differ between $\mathbf{x}$ and $\tilde{\mathbf{y}}$, and there are $k-j+1$ variables that differ between $\tilde{\mathbf{x}}$ and $\mathbf{y}$. Therefore, by the induction hypothesis, both terms are greater than or equal to 1. □

By computing the second derivative of the Gaussian density function in information form, this theorem directly implies that a multivariate Gaussian distribution with information matrix $\mathbf{J}$ is MTP_2 if and only if

$$\mathbf{J}_{ij} \leq 0 \text{ for all } i \neq j.$$

Let us pause for a moment and think: what does it mean for $\mathbf{J}_{ij}$ to be less than or equal to zero and how does this relate to positive dependence? Specifically, for $i \geq j$, what is the conditional distribution of $\mathsf{x}_i, \mathsf{x}_j$ conditioned on everything else? From the lectures on Gaussian graphical models, we know:

$$\operatorname{cov}(\mathsf{x}_i, \mathsf{x}_j \mid \mathbf{x}_{\mathcal{V} \setminus \{i,j\}}) = \begin{bmatrix} \mathbf{J}_{ii} & \mathbf{J}_{ij} \\ \mathbf{J}_{ji} & \mathbf{J}_{jj} \end{bmatrix}^{-1} = \frac{1}{\mathbf{J}_{ii}\mathbf{J}_{jj} - \mathbf{J}_{ij}^2} \begin{bmatrix} \mathbf{J}_{jj} & -\mathbf{J}_{ij} \\ -\mathbf{J}_{ij} & \mathbf{J}_{ii} \end{bmatrix}.$$

So, $\mathbf{J}_{ij} \leq 0$ means that the conditional covariance between x_i and x_j given all other nodes is greater than or equal to zero. Hence, having $\mathbf{J}_{ij} \leq 0$ corresponds to a form of positive dependence.

An interesting linear algebra fact is that if a positive definite matrix (hence invertible) has all off-diagonal entries non-positive, then its inverse has all entries non-negative. Hence, since the information matrix $\mathbf{J}$ is positive definite and $\mathbf{J}_{ij} \leq 0$ for all off-diagonal entries, then the covariance matrix $\mathbf{\Lambda} = \mathbf{J}^{-1}$ satisfies $\mathbf{\Lambda}_{ij} \geq 0$ for all i, j. This means that not only the conditional correlations are positive, but also all the (unconditional) correlations are positive. In fact, no matter which variables we condition on, we get positive dependence between the random variables. This shows that MTP_2 is a very strong form of positive dependence.

Example 3. A 2016 data of monthly correlations of global stock markets (taken from InverstmentFrontier.com) read

$$\mathbf{S} = \begin{array}{c} \begin{array}{ccccc} \text{Nasdaq} & \text{Canada} & \text{Europe} & \text{UK} & \text{Australia} \end{array} \\ \begin{bmatrix} 1.000 & 0.606 & 0.731 & 0.618 & 0.613 \\ 0.606 & 1.000 & 0.550 & 0.661 & 0.598 \\ 0.731 & 0.550 & 1.000 & 0.644 & 0.569 \\ 0.618 & 0.661 & 0.644 & 1.000 & 0.615 \\ 0.613 & 0.598 & 0.569 & 0.615 & 1.000 \end{bmatrix} \end{array} \begin{array}{c} \\ \text{Nasdaq} \\ \text{Canada} \\ \text{Europe} \\ \text{UK} \\ \text{Australia} \end{array}$$

The inverse of the sample covariance matrix is evaluated to

$$\mathbf{S}^{-1} = \begin{array}{c} \begin{array}{ccccc} \text{Nasdaq} & \text{Canada} & \text{Europe} & \text{UK} & \text{Australia} \end{array} \\ \begin{bmatrix} 2.629 & -0.480 & -1.249 & -0.202 & -0.490 \\ -0.480 & 2.109 & -0.039 & -0.790 & -0.459 \\ -1.249 & -0.039 & 2.491 & -0.675 & -0.213 \\ -0.202 & -0.790 & -0.675 & 2.378 & -0.482 \\ -0.490 & -0.459 & -0.213 & -0.482 & 1.992 \end{bmatrix} \end{array} \begin{array}{c} \\ \text{Nasdaq} \\ \text{Canada} \\ \text{Europe} \\ \text{UK} \\ \text{Australia} \end{array}$$

so the sample distribution is MTP_2. It is quite remarkable that MTP_2 correlation matrices can be found in practice, since if you were to sample correlation matrices uniformly at random the probability of it being MTP_2 is less than 10^{-6}.

9.3 Properties of MTP_2 distributions

Next, we will see some of the nice properties of MTP_2 distributions. As we have seen for Gaussian distributions and for undirected graphical models, MTP_2 distributions are closed under marginalization and conditioning.

Theorem 2 ([2]). *If a joint distribution f is MTP_2, then*

(i) any conditional distribution is MTP_2,

(ii) any marginal distribution is MTP_2.

While the proof that any conditional of an MTP_2 distribution is MTP_2 follows directly from the defintion of MTP_2, the proof for marginalization was given in [2] and is based on the fact that any MTP_2 functions f and g satisfy

(a) their product $f(\mathbf{x})g(\mathbf{x})$ is MTP_2,

(b) their composition $h(\mathbf{x}, \mathbf{z}) \triangleq \int f(\mathbf{x}, \mathbf{y})g(\mathbf{y}, \mathbf{z})d\mathbf{y}$ is MTP_2.

This shows that MTP_2 distributions are nice, in the sense that whenever you have an MTP_2 distribution, you are free to do some conditioning or marginalization without losing the MTP_2 property.

The reason for discussing MTP_2 distributions in this lecture is to provide another example of a constraint that interfaces nicely with graphical models. For the remainder of this lecture we will therefore discuss properties of MTP_2 distributions that are related to conditional independence implications.

We start with a prominent theorem for positively associated distributions, which hence of course also holds for MTP_2 distributions. Namely, the following theorem shows that for positively associated distributions zero covariance implies independence. Note that the $\Longrightarrow$ part of the theorem is true for any distribution, but the $\Longleftarrow$ part is special. We have seen that this holds for Gaussian distributions, and the following theorem shows that this is also true for positively associated distributions.

Theorem 3 ([5]). *If* $\mathbf{x}$ *is positively associated and* $\mathcal{A}, \mathcal{B} \subset \mathcal{V}$ *are disjoint, then*

$$\mathbf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{x}_{\mathcal{B}} \iff \operatorname{cov}(\mathbf{x}_u, \mathbf{x}_v) \text{ for all } u \in \mathcal{A}, v \in \mathcal{B}.$$

In fact, it was recently shown that MTP_2 distributions satisfy very strong conditional independence implications that do not generally hold for positively associated distributions or even Gaussian distributions, but are very natural if you think of conditional independence constraints as implied by separation statements in an undirected graph. In particular, in an undirected graph if two sets of vertices $\mathcal{A}$ and $\mathcal{B}$ are separated by a set of vertices $\mathcal{C}$, then $\mathcal{A}$ and $\mathcal{B}$ remain separated even when adding nodes to $\mathcal{C}$. This property is known as *upward stability* and it turns out that conditional independence relations in MTP_2 distributions satisfy this property. In addition, in an undirected graph it holds that if two nodes i, j are separated, then they are in two different connected components and hence any third node k is either in the connected component of j and hence separated from i or it is in the connected component of i and hence separated from j or it is in a different connected component all together and hence separated from both nodes i and j. This property is known as *singleton transitivity* and it turns out that conditional independence relations in MTP_2 distributions satisfy this property. This is shown in the following theorem.

Theorem 4 ([6]). *If* $\mathbf{x}$ *is* MTP_2*, then its independence model satisfies:*

(i) Upward stability:

$$\mathbf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{x}_{\mathcal{B}} \mid \mathbf{x}_{\mathcal{C}} \implies \mathbf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{x}_{\mathcal{B}} \mid \mathbf{x}_{\mathcal{C} \cup \{k\}}.$$

(ii) Singleton transitivity:

$$\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \implies \mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_k \text{ or } \mathbf{x}_j \perp\!\!\!\perp \mathbf{x}_k.$$

These conditional independence implications suggest a very strong connection between MTP_2 distributions and undirected graphical models. In fact, it turns out that for any strictly positive MTP_2 distribution there exists an undirected graph so that the conditional independence relations in the distribution exactly correspond to the separation statements in the undirected graph. In other words, for any strictly positive MTP_2 distribution there exists an undirected graph which is a perfect map with respect to the distribution. This is a strong statement; in general, there only exists a minimal I-map that satisfies $\mathcal{I}(\mathcal{G}) \subsetneq \mathcal{I}(p)$, and we have seen examples of distributions that do not have undirected graphs that are P-maps for the distributions.

Theorem 5 ([6]). *Let* $p(\mathbf{x}) > 0$ *be* MTP_2*. Then there exists an undirected graph* $\mathcal{G}$ *such that* $\mathcal{I}(p) = \mathcal{I}(\mathcal{G})$*. This graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *is given by:*

$$(i, j) \notin \mathcal{E} \iff \mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{x}_{\mathcal{V} \setminus \{i,j\}}.$$

Proof. Recall that the pairwise Markov property implies the global Markov property for strictly positive distributions. Therefore, any conditional independence relation that holds in $\mathcal{G}$ also holds in p. It is left to show that any conditional independence relation that does *not* hold in $\mathcal{G}$ also does not hold in p.

Consider disjoint sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{V}$ such that $\mathcal{A}$ and $\mathcal{B}$ are not separated in $\mathcal{G}$ when $\mathcal{C}$ is removed from $\mathcal{G}$. We aim to show that $\mathbf{x}_\mathcal{A} \not\perp\!\!\!\perp \mathbf{x}_\mathcal{B} \mid \mathbf{x}_\mathcal{C}$. Then, there are nodes $u \in \mathcal{A}$ and $v \in \mathcal{B}$ and a path $u = v_1, v_2, \ldots, v_r = v$ such that $v_i \notin \mathcal{C}$ for all $i = 1, \ldots, r$. For every edge (v_i, v_{i+1}) we have $\mathbf{x}_{v_i} \not\perp\!\!\!\perp \mathbf{x}_{v_{i+1}} \mid \mathbf{x}_{\mathcal{V}\setminus\{i,i+1\}}$, and hence $\mathbf{x}_{v_i} \not\perp\!\!\!\perp \mathbf{x}_{v_{i+1}} \mid \mathbf{x}_\mathcal{C}$, by upward stability. Then, by singleton transitivity we have $\mathbf{x}_u \not\perp\!\!\!\perp \mathbf{x}_v \mid \mathbf{x}_\mathcal{C}$. □

In fact, the following related statement also holds: For any undirected graph $\mathcal{G}$ there exists an MTP_2 distribution such that $\mathcal{I}(p) = \mathcal{I}(\mathcal{G})$. This can for example be attained by an MTP_2 Gaussian distribution. While these results prove that MTP_2 distributions interact nicely with undirected graphical models, they also show that MTP_2 distributions do not interact so nicely with directed graphical models. In particular, given a DAG $\mathcal{G}$, in general there does not exist an MTP_2 distribution p such that $\mathcal{I}(p) = \mathcal{I}(\mathcal{G})$. Such a P-map only exists for directed graphs that have an undirected P-map. In particular, the 3-node DAG with edges $1 \to 3$ and $2 \to 3$ cannot be a P-map of an MTP_2 distribution.

References

[1] Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[2] Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

[3] JHB Kemperman. On the FKG-inequality for measures on a partially ordered space. In *Indagationes Mathematicae (Proceedings)*, volume 80, pages 313–331. North-Holland, 1977.

[4] Samuel Karlin and Yosef Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear algebra and its applications*, 52:419–438, 1983.

[5] Joel L Lebowitz. Bounds on the correlations and analyticity properties of ferromagnetic Ising spin systems. *Communications in Mathematical Physics*, 28(4):313–321, 1972.

[6] Shaun Fallat, Steffen Lauritzen, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, Piotr Zwiernik, et al. Total positivity in Markov structures. *The Annals of Statistics*, 45(3):1152–1184, 2017.