The Annals of Statistics
2021, Vol. 49, No. 3, 1436–1459
https://doi.org/10.1214/20-AOS2007

TOTAL POSITIVITY IN EXPONENTIAL FAMILIES WITH APPLICATION TO BINARY VARIABLES

BY STEFFEN LAURITZEN[1], CAROLINE UHLER[2] AND PIOTR ZWIERNIK[3]

[1]*Department of Mathematical Sciences, University of Copenhagen, lauritzen@math.ku.dk*

[2]*Laboratory for Information and Decision Systems, and Institute for Data, Systems, and Society Massachusetts Institute of Technology, cuhler@mit.edu*

[3]*Department of Economics and Business, Universitat Pompeu Fabra, piotr.zwiernik@upf.edu*

We study exponential families of distributions that are multivariate totally positive of order 2 (MTP_2), show that these are convex exponential families and derive conditions for existence of the MLE. Quadratic exponential familes of MTP_2 distributions contain attractive Gaussian graphical models and ferromagnetic Ising models as special examples. We show that these are defined by intersecting the space of canonical parameters with a polyhedral cone whose faces correspond to conditional independence relations. Hence MTP_2 serves as an implicit regularizer for quadratic exponential families and leads to sparsity in the estimated graphical model. We prove that the maximum likelihood estimator (MLE) in an MTP_2 binary exponential family exists if and only if both of the sign patterns $(1, -1)$ and $(-1, 1)$ are represented in the sample for every pair of variables; in particular, this implies that the MLE may exist with $n = d$ observations, in stark contrast to unrestricted binary exponential families where 2^d observations are required. Finally, we provide a novel and globally convergent algorithm for computing the MLE for MTP_2 Ising models similar to iterative proportional scaling and apply it to the analysis of data from two psychological disorders.

1. Introduction and motivation. This paper discusses exponential families and, in particular, binary graphical models with a special form of positive dependence. Total positivity is a strong form of positive dependence that has become an important concept in modern statistics; see, for example, [14, 22]. This property (also called the MTP_2 property) appeared in the study of stochastic orderings, asymptotic statistics and in statistical physics [19, 30]. Families of distributions with this property lead to many computational advantages [8, 16, 31] and they are a convenient shape constraint in nonparametric statistics [32]. They also became a useful tool in modelling with latent variables; see [9] for an overview. In particular, in [4] the MTP_2 property explicitly appeared in the description of the binary latent class model.

In the Gaussian setting, the MTP_2 property was shown to simplify inference [5, 27]. In this case, the MTP_2 property is equivalent to the covariance matrix being an inverse M-matrix, which is a linear constraint on the concentration matrix. This led Slawski and Hein [33] to propose efficient learning procedures based on convex optimization; see also [11, 17, 25]. The present paper develops similar results for exponential families with special emphasis on models for binary variables, including ferromagnetic Ising models. Our main results are the following:

- We show in Section 3 that the MTP_2 property is given by a convex constraint in an exponential family and use convex optimization theory to derive necessary and sufficient conditions ensuring than an estimate maximizes the likelihood. For a quadratic exponential

Received February 2020; revised May 2020.
MSC2020 subject classifications. Primary 60E15, 62H99; secondary 15B48.
Key words and phrases. Graphical models, Ising model, log-supermodular distributions, positive dependence, exponential families.

family, including the Ising model for binary variables, the KKT conditions yield sparsity in the associated matrix for interaction potentials.

- We show in Section 4 that the KKT conditions ensure context-dependent conditional independence restrictions and that for binary variables the MLE exists under MTP_2 if and only if both of the sign patterns $(1, -1)$ and $(-1, 1)$ are represented in the sample for every pair of variables. This ensures the minimal sample size for the MLE to exist be of order d rather than 2^d where d is the number of variables considered.
- We show—also in Section 4—that adding conditional independence assumptions by further assuming a graphical model, reduces this condition to hold for pairs of vertices ij that are neighbors in the graph, reducing the order of the minimal sample size to be the maximal clique size of the graph.
- We show—also in Section 4—that for symmetric binary MTP_2 distributions, including ferromagnetic Ising models with no external field, presence of just one of the sign patterns $(1, -1)$ and $(-1, 1)$ for every pair ensures existence of the MLE;
- We develop—in Section 5—a novel IPS type algorithm for calculating the MLE in a ferromagnetic Ising model that is shown to be globally convergent.

The remainder of this paper is structured as follows: In Section 2, we formally introduce MTP_2 distributions and associated notation.

In Section 6, we apply our results to the analysis of two psychological disorders, showing that the resulting MTP_2 graphical model is highly interpretable and consistent with domain knowledge.

2. Preliminaries. Let $V = \{1, \dots, d\}$ be a finite set and let $X = (X_v, v \in V)$ be random variables with labels in V. 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.

ASSUMPTION 1. All distributions are assumed to have densities with respect to the product measure $\mu = \bigotimes_{v \in V} \mu_v$, referred to as the *base measure*, where μ_v is the counting measure if $\mathcal{X}_v$ is discrete, and μ_v is the Lebesgue measure giving length 1 to the unit interval if $\mathcal{X}_v$ is an open interval.

We note that any other equivalent product measure can be used as base measure without affecting the MTP_2 property as defined below.

A function f on $\mathcal{X}$ is said to be *multivariate totally positive of order* 2 (MTP_2) if

$$f(x) f(y) \le f(x \wedge y) f(x \vee y) \quad \text{for all } x, y \in \mathcal{X}, \tag{2.1}$$

where $x \wedge y$ and $x \vee y$ denote the elementwise minimum and maximum, that is,

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

These inequalities are nontrivial only if $x, y \in \mathcal{X}$ are *not comparable*, that is, neither $x \le y$ nor $x \ge y$. For $d = 2$, a function that is MTP_2 is simply called *totally positive* [22]. We say that X or the distribution of X is MTP_2 if its density function p is MTP_2.

For strictly positive distributions, MTP_2 can be verified by checking that (2.1) holds for $x, y \in \mathcal{X}$ that are not comparable and differ in exactly two coordinates; cf. [22], Proposition 2.1. We call such pairs *elementary* and denote the set of all elementary pairs by $\mathcal{E} \subset \mathcal{X} \times \mathcal{X}$. For more details on MTP_2 distributions, see [22] and [18].

3. Totally positive exponential families. We first consider MTP_2 for exponential families and show that maximum likelihood estimation for exponential families under MTP_2 leads to a convex optimization problem. We then discuss conditions for the existence of the MLE and finally specialize these results to quadratic exponential families, which include as prominent examples the Gaussian distribution and the Ising model.

3.1. *Convexity of totally positive exponential families.* Consider an exponential family with density $p(x;\theta)$ satisfying

$$\log p(x;\theta) = \langle \theta, T(x)\rangle - A(\theta) + g(x), \tag{3.1}$$

with sample space $\mathcal{X}$, sufficient statistics $T : \mathcal{X} \to \mathbb{R}^k$ and base measure μ. Assume that the family is *minimally represented*, that is, that $\langle \lambda, T(X)\rangle + b = 0$ almost surely implies $\lambda = 0$, and that the family is *regular* so that the space of canonical parameters

$$\mathcal{K} = \{\theta \in \mathbb{R}^k : A(\theta) < \infty\}$$

is an open convex set.

ASSUMPTION 2. Throughout, we assume that there exists θ_0 such that $p(x;\theta_0)$ is a product distribution, or equivalently,

$$p(x \vee y;\theta_0)p(x \wedge y;\theta_0) = p(x;\theta_0)p(y;\theta_0) \quad \text{for all } x, y \in \mathcal{X}. \tag{3.2}$$

Since every distribution in an exponential family can act as the base distribution, we can then pick $p(x;\theta_0)$ as the base measure. It then holds that

$$g(x \vee y) + g(x \wedge y) - g(x) - g(y) = 0.$$

We say that such an exponential family *has a product base*.

All exponential families that contain a full independence distribution admit a product base. This includes all models discussed in this article and in particular Gaussian graphical models and log-linear models.

For an exponential family of the form (3.1) and any two $x, y \in \mathcal{X}$ we define

$$\Delta(x, y;\theta) := \log\left(\frac{p(x \vee y;\theta)p(x \wedge y;\theta)}{p(x;\theta)p(y;\theta)}\right).$$

The density $p(x;\theta)$ is MTP_2 if and only if $\Delta(x, y;\theta) \geq 0$ for all elementary pairs in $\mathcal{E}$. For exponential families with a product base, it holds that

$$\Delta(x, y;\theta) = \langle \theta, T(x \wedge y) + T(x \vee y) - T(x) - T(y)\rangle,$$

which is an *affine* function in θ.

DEFINITION 3.1. The set $\mathcal{K}_2 \subset \mathcal{K}$ of totally positive canonical parameters is the subset of canonical parameters for which the density $p(x;\theta)$ is MTP_2.

Since $\mathcal{K}_2$ is given by the linear inequalities $\Delta(x, y;\theta) \geq 0$ for all $x, y \in \mathcal{X}$, we immediately get the following result.

THEOREM 3.2. *The $\mathcal{K}_2$ of totally positive canonical parameters is a convex set that is relatively closed in $\mathcal{K}$.*

We note that this result holds also for exponential families without a product base. However, in that case the set of MTP_2 canonical parameters $\mathcal{K}_2$ may be empty.

In [25], we considered the Gaussian setting and showed that $\mathcal{K}_2$ is a convex cone. By essentially the same argument, this extends to discrete Gaussian distributions over $\mathcal{X} = \mathbb{Z}^d$, which were introduced in [1]. More generally, we obtain the following result.

PROPOSITION 3.3. *The set $\mathcal{K}_2$ is obtained by intersecting $\mathcal{K}$ with a closed convex cone $\mathcal{C} \subseteq \mathbb{R}^k$, whose dual cone is the closure of the cone generated by the set*

$$\{T(x \wedge y) + T(x \vee y) - T(x) - T(y) : x, y \in \mathcal{E}\}.$$

PROOF. The set of inequalities $\Delta(x, y; \theta) \geq 0$, one for each elementary pair $x, y \in \mathcal{E}$, defines a convex cone in $\theta \in \mathbb{R}^k$. We have $\langle \theta, T(x \wedge y) + T(x \vee y) - T(x) - T(y) \rangle \geq 0$ for all $x, y \in \mathcal{E}$ if and only if $\langle \theta, v \rangle \geq 0$ for all v in the cone generated by the set $\{T(x \wedge y) + T(x \vee y) - T(x) - T(y) : x, y \in \mathcal{E}\}$; denote this cone by $\mathcal{C}^*$. This shows that $\mathcal{C} = (\mathcal{C}^*)^\vee$ and so $\mathcal{C}^\vee = (\mathcal{C}^*)^{\vee\vee}$. The latter is equal to the closure of $\mathcal{C}^*$ by the standard theory of convex cones; see, for example, [13], Section 2.6.1. □

REMARK 3.4. When $\mathcal{X}$ is finite, that is, for log-linear models, Proposition 3.3 implies that $\mathcal{C}$ is polyhedral. Since $\mathcal{C}$ is polyhedral also in the Gaussian setting, finiteness of $\mathcal{X}$ is not a necessary condition. In fact, we will show in Proposition 3.6 that $\mathcal{C}$ is polyhedral for any quadratic exponential family. When $\mathcal{C}$ is polyhedral, then every face of $\mathcal{C}$ intersected with $\mathcal{K}$ corresponds to the MTP_2 distributions in an exponential subfamily.

3.2. *The MLE and its existence.* An important consequence of Theorem 3.2 is that any MTP_2 exponential family is a convex exponential family and thus the maximum likelihood estimator (MLE), if it exists, is uniquely defined; see [7], Section 9.4.

Let $U = \{x^1, \ldots, x^n\}$ denote a sample of size n and let $\bar{T} := \frac{1}{n}\sum_i T(x^i)$ be the average of the corresponding sufficient statistics. Let $\mathcal{S}$ denote the interior of $\operatorname{conv}(\operatorname{supp}(\mu \circ T^{-1}))$, the convex support of the sufficient statistics. Then by the general theory of exponential families [7], the MLE $\hat{\theta}$ exists if and only if $\bar{T}$ lies in $\mathcal{S}$, in which case it is *uniquely* defined by

$$\nabla A(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}[T(X)] = \bar{T}.$$

The following theorem extends this result to a characterization of existence of the MLE for the subfamily of MTP_2 distributions. By Proposition 3.3, there exists a closed convex cone $\mathcal{C}$ such that the space of all MTP_2 canonical parameters is given by $\mathcal{K}_2 = \mathcal{K} \cap \mathcal{C}$. We define

$$\mathcal{S}_2 := \mathcal{S} - \mathcal{C}^\vee$$

as the Minkowski sum of $\mathcal{S}$ with the dual of $-\mathcal{C}$; cf. Proposition 3.3.

THEOREM 3.5. *Let $p(x; \theta)$ be a minimally represented regular exponential family. Then the MLE $\hat{\theta}$ based on $\bar{T}$ exists in the* MTP_2 *submodel if and only if $\bar{T} \in \mathcal{S}_2$, in which case $\hat{\theta}$ is uniquely defined by*:

(a) *primal feasibility*: $\hat{\theta} \in \mathcal{K}_2$,
(b) *dual feasibility*: $\hat{\sigma} := \nabla A(\hat{\theta}) \in \mathcal{S}$ *with* $\hat{\sigma} - \bar{T} \in \mathcal{C}^\vee$,
(c) *complementary slackness*: $\langle \hat{\theta}, \hat{\sigma} - \bar{T} \rangle = 0$.

PROOF. The maximum likelihood estimation problem can be formulated as the following optimization problem:

$$\begin{aligned} &\underset{\theta\in\mathcal{K}}{\text{maximize}} \quad \langle\theta,\bar{T}\rangle - A(\theta)\\ &\text{subject to } \theta\in\mathcal{C}. \end{aligned}$$

This is a convex optimization problem, since $A(\theta)$ is convex on $\mathcal{K}$. The Lagrangian is

$$\mathcal{L}(\theta,\lambda) = \langle\theta,\bar{T}\rangle - A(\theta) + \langle\theta,\lambda\rangle,$$

where $\lambda\in\mathcal{C}^\vee$. Let A^* denote the conjugate dual of A with domain $\mathcal{S}$. Then

$$\max_{\theta\in\mathcal{K}} \mathcal{L}(\theta,\lambda) = A^*(\bar{T}+\lambda),$$

and hence the dual optimization problem is given by

$$\begin{aligned} &\underset{\sigma\in\mathcal{S}}{\text{minimize}} \quad A^*(\sigma)\\ &\text{subject to } \sigma-\bar{T}\in\mathcal{C}^\vee. \end{aligned}$$

The MLE exists if and only if the primal and dual problems are feasible. The primal problem is feasible by the assumption $\mathcal{K}_2\neq\varnothing$. The dual problem is feasible if and only if $\bar{T}\in\mathcal{S}_2$. The characterization of the MLE then follows from the KKT conditions. □

As in the Gaussian case, complimentary slackness imposes sparsity in the MLE $\hat{\theta}$. This property makes MTP_2 exponential families potentially useful in high dimensional contexts. Before we discuss this in further detail, we shall consider the case of a quadratic exponential family, including the Gaussian case and Ising models.

3.3. *Quadratic exponential families.* The density function of a *quadratic exponential family* is of the form

$$p(x;h,J) = \exp\bigl(h^Tx + x^TJx/2 - A(h,J)\bigr), \tag{3.3}$$

with $h\in\mathbb{R}^d$ and $J\in\mathbb{S}^d$, where $\mathbb{S}^d$ is the set of symmetric matrices in $\mathbb{R}^{d\times d}$ so here the canonical parameter space is $\mathcal{K}=\mathbb{R}^d\times\mathbb{S}^d$. Important examples of such exponential families in the discrete setting are Ising models, which we discuss in more detail in Section 5, and Gaussian graphical models in the continuous setting. Note that in the binary setting we require $J_{ii}=0$ in order to obtain a minimally represented exponential family. We start by showing that $\mathcal{C}$ is a polyhedral cone for any quadratic exponential family.

PROPOSITION 3.6. *The subfamily of* MTP_2 *distributions in a quadratic exponential family is obtained by intersecting* $\mathcal{K}$ *with a polyhedral cone* $\mathcal{C}$, *namely the cone* $\mathbb{S}^d_+ = \{J\in\mathbb{S}^d \mid J_{ij}\geq 0$ *for all* $i\neq j\}$.

PROOF. By [18], Theorem 7.5, a quadratic exponential family is MTP_2 if and only if $\exp(J_{ij}x_ix_j)$ is MTP_2 for all $i\neq j$. This is the case if and only if for every x, y that differ in two coordinates i, j with $x_i<y_i$ and $x_j>y_j$, it holds that

$$J_{ij}(y_i-x_i)(x_j-y_j)\geq 0,$$

or equivalently $J_{ij}\geq 0$. This completes the proof. □

We denote the mean parameters by $\mu := \mathbb{E}_\theta X$ and $\Xi := \mathbb{E}_\theta XX^T$. Then (μ,Ξ) can be transformed to (μ,Σ), where $\Sigma=\Xi-\mu\mu^T$ is the covariance matrix of X. Note that then

$$\mathcal{C} = \bigl\{(h,J)\in\mathbb{R}^d\times\mathbb{S}^d_+ : J_{ij}\geq 0 \text{ for } i\neq j\bigr\}.$$

Each facet of $\mathcal{C}$ corresponds to one of the J_{ij}'s being zero; cf. Remark 3.4. Equivalently, by the Hammersley–Clifford theorem, each facet consists of members in the MTP_2 exponential family that satisfy the conditional independence relation $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$. The dual cone of $\mathcal{C}$ is given by

$$\mathcal{C}^\vee = \{(0, \Xi) \in \mathbb{R}^d \times \mathbb{S}^d : \Xi_{ij} \geq 0 \text{ for } i \neq j, \text{ and } \Xi_{ii} = 0 \text{ for all } i\}. \tag{3.4}$$

Let $U = \{x^1, \ldots, x^n\}$ as before be a sample of size n and let $\bar{x} = \frac{1}{n}\sum_i x^i$ and $M = \frac{1}{n}\sum_i x^i (x^i)^T$ be the corresponding sample averages. Let $S = M - \bar{x}\bar{x}^T$ denote the sample covariance matrix. By standard exponential family theory, the MLE in the quadratic exponential family (3.3) corresponds to the unique distribution in the family which matches the sample averages, that is, $(\hat{\mu}, \hat{\Xi}) = (\bar{x}, M)$, or equivalently, $(\hat{\mu}, \hat{\Sigma}) = (\bar{x}, S)$. By adding the MTP_2 constraint, the situation changes somewhat. As a direct corollary to Theorem 3.5 we obtain the following result regarding the MLE in an MTP_2 quadratic exponential family.

COROLLARY 3.7. *Let $p(x; h, J)$ be a minimal regular quadratic exponential family. Let $\bar{x}$ and S be the sample mean and covariance matrix. Then the corresponding MLE $(\hat{h}, \hat{J}) \in \mathcal{K}$ with $(\hat{\mu}, \hat{\Xi}) := \nabla A(\hat{h}, \hat{J})$ and $\hat{\Sigma} := \hat{\Xi} - \hat{\mu}\hat{\mu}^T$, is uniquely defined by*:

(i) $\hat{J}_{ij} \geq 0$ *for* $i \neq j$,
(ii) $\hat{\mu} = \bar{x}$, $\hat{\Sigma}_{ii} = S_{ii}$, *and* $\hat{\Sigma}_{ij} \geq S_{ij}$ *for* $i \neq j$,
(iii) $(\hat{\Sigma}_{ij} - S_{ij})\hat{J}_{ij} = 0$ *for all* $i \neq j$.

PROOF. The conditions of Theorem 3.5 translate precisely to (i), (ii), (iii), namely the primal feasibility condition is derived in Proposition 3.6, the dual feasibility condition follows from (3.4), and the complementary slackness condition follows from the fact that the inner product between dual cones is zero if and only if each summand is zero. □

REMARK 3.8. In quadratic exponential families the condition $\overline{T} \in \mathcal{S}_2 = \mathcal{S} - \mathcal{C}^\vee$, that assures existence of the MLE, translates to the condition (ii) in Corollary 3.7. This condition can again be expressed more explicitly in terms of the observations: in the Gaussian case this becomes equivalent to all correlations being numerically less than one ([25]), and we derive the explicit conditions for our cases in Theorem 4.5, Corollary 4.6 and Theorem 4.11.

REMARK 3.9. Note also that in the binary case, where we have $J_{ii} = 0$ and $\Xi_{ii} = 1$ for all i, the condition (ii) reduces to $\hat{\mu} = \bar{x}$, and $\hat{\Sigma}_{ij} \geq S_{ij}$ for $i \neq j$.

The specialization of this result to Gaussian graphical models was discussed in [25]. Note that the MTP_2 constraint induces sparsity in the MLE $\hat{J}$ through the complementary slackness constraint (iii). For example, if $S_{ij} < 0$, then complementary slackness implies that $\hat{J}_{ij} = 0$ simply because in an MTP_2 distribution all covariances are positive. The sparsity pattern of $\hat{J}$ defines a face $\mathcal{F}$ of the polyhedral cone $\mathcal{C}$. As in the Gaussian setting [25], Corollary 2.4, the MTP_2 MLE $\hat{J}$ is the MLE of the quadratic exponential family without the MTP_2 constraint restricted to the face $\mathcal{F}$. This is stated formally in Corollary 3.10 and illustrated in Example 5.3 below.

COROLLARY 3.10. *Let $\hat{J}$ denote the MLE in a quadratic exponential family under* MTP_2. *Let $\mathcal{F} = \{(i, j) \in V \times V \mid \hat{J}_{ij} = 0\}$. Then $\hat{J}$ equals the maximum likelihood estimate in the quadratic exponential family without the* MTP_2 *constraint under the linear constraints $J_{ij} = 0$ for all $(i, j) \in \mathcal{F}$.*

PROOF. This follows since the unique MLE in this quadratic exponential family is given by the equations (ii) and (iii) in Corollary 3.7 above. □

In [25], it was shown that the MLE existed in the Gaussian case if and only if the empirical covariance matrix satisfied $S_{ij} < \sqrt{S_{ii}S_{jj}}$ by constructing an ultrametric matrix Z from S that was both primary and dually feasible. The argument used in [25] does not apply here as the primary feasibility of Z is not always guaranteed. Indeed, we shall see that the condition is necessary here but not sufficient; see Theorem 4.5 and Corollary 4.6 below. The situation in a general exponential family can be quite different from the Gaussian case as shown in the following example.

EXAMPLE 3.11. The auto-Poisson family considered in [10], Section 4.2.4, is a quadratic exponential family with product base. It consists of distributions of the form

$$p(x; h, J) \propto \exp\left(\sum_{i=1}^{d}\big(h_i x_i - \log(x_i!)\big) + x^T Jx/2\right) \quad x \in \{0, 1, 2, \ldots\}^d.$$

The right-hand side sums to a finite number if and only if $J_{ij} \leq 0$ for all i, j. The subset of MTP_2 distributions within this family is then given by the product of independent Poisson distributions, that is, $J_{ij} = 0$ for all i, j. Of course, for a finite state-space, no such problem occurs.

4. Totally positive binary distributions. For the remainder of this paper, we focus on binary distributions, that is, distributions over the sample space $\mathcal{X} = \{-1, 1\}^d$. To simplify notation, we often use the following bijection between $\mathcal{X}$ and the set $\mathbf{B}_d$ of all subsets of $\{1, \ldots, d\}$, namely an element $x \in \mathcal{X}$ maps to the subset of all $i \in \{1, \ldots, d\}$ for which $x_i = 1$. For example, in the case $d = 3$ the point $x = (1, 1, -1)$ maps to the subset $\{1, 2\}$ and $(-1, -1, -1)$ to the empty set. Note that $\mathcal{X}$ and $\mathbf{B}_d$ are also isomorphic as lattices because the min-max operators $\wedge, \vee$ on $\mathcal{X}$ correspond to the set operations $\cap, \cup$ in $\mathbf{B}_d$.

Building on the results from Section 3, in the following we provide conditions for existence of the MLE in MTP_2 binary exponential families. In particular, we study the KKT conditions for this setting and develop conditions for existence of the MLE in the special case of binary distributions that factorize according to a graph (such as Ising models) and symmetric binary distributions where $p(x) = p(-x)$ (such as Ising models with no external field). Ising models will be discussed in detail in Section 5.

4.1. *Binary distributions as exponential families.* We now recall the representation of strictly positive binary distributions as an exponential family. Define $\lambda(x) := \log p(x)$ for $x \in \mathcal{X} = \{-1, 1\}$. To write the exponential representation of this family of distributions, we consider the space $\mathbb{R}^{\mathcal{X}}$ of dimension 2^d equipped with the inner product

$$\langle \theta, \sigma \rangle := \sum_{x \in \mathcal{X}} \theta(x)\sigma(x).$$

For $x \in \mathcal{X}$, define a vector $T(x) \in \{0, 1\}^{\mathcal{X}}$ such that $T(x)_y = 1$ if $x = y$ and it is zero otherwise. The set of binary distributions forms a regular exponential family which is minimally represented with canonical parameters $\theta(x) = \lambda(x) - \lambda(-\mathbf{1})$ for $x \neq -\mathbf{1}$. Denote by θ the vector of all $\theta(x)$ for $x \in \mathcal{X}$ and observe that $\theta(-\mathbf{1}) = 0$. Then

$$p(x) = \exp\big(\langle \theta, T(x) \rangle - A(\theta)\big),$$

where $A(\theta) = \log[\langle \mathbf{1}, \exp(\theta)\rangle]$. The space of canonical parameters is simply the $2^d - 1$ dimensional real vector space $\mathbb{R}^{\mathcal{X}'}$ where $\mathcal{X}' = \mathcal{X} \setminus \{-\mathbf{1}\}$. The interior of the convex support of the sufficient statistics is given by the set

$$\mathcal{S} = \Big\{ p \in \mathbb{R}^{\mathcal{X}'} : \, p(x) > 0 \text{ for all } x \in \mathcal{X}' \text{ and } \sum_{x \in \mathcal{X}'} p(x) < 1 \Big\},$$

which we identify with the interior of the probability simplex, namely

$$\mathcal{S}_1 = \Big\{ p : \, p(x) > 0 \text{ for all } x \in \mathcal{X} \text{ and } \sum_{x \in \mathcal{X}} p(x) = 1 \Big\}.$$

REMARK 4.1. The constraints on the space of canonical parameters $\mathcal{K}$ defining binary MTP_2 distributions are

$$\theta(x \wedge y) + \theta(x \vee y) - \theta(x) - \theta(y) \geq 0 \tag{4.1}$$

for all elementary pairs $x, y \in \mathcal{X}$. We recall that a pair x, y is elementary if there exist a subset $A \subset V$ and $i, j \in V \setminus A$ such that x corresponds to $A \cup \{i\}$ and y corresponds to $A \cup \{j\}$. The number of such pairs is $\binom{d}{2} 2^{d-2}$. Another way to phrase (4.1) is that θ is a supermodular set-function that satisfies the normalizing condition $\theta(-\mathbf{1}) = 0$; cf. [6].

4.2. *KKT conditions and conditional independence.* In this section, we study how the KKT conditions of Theorem 3.5 induce sparsity in the general binary setting, in the form of context-specific conditional independence constraints. To do this, we introduce some notation. Following Studený [34], we call the elements in $\mathbb{Z}^{\mathcal{X}}$ *imsets*. An important example of an imset is $T(x) \in \{0, 1\}^{\mathcal{X}}$ defined earlier. The imset

$$u_{x,y} := T(x \wedge y) + T(x \vee y) - T(x) - T(y)$$

is called a *semielementary imset*. If x, y form an elementary pair then $u_{x,y}$ is called an *elementary imset*. If this pair is associated to sets $A \cup \{i\}$ and $A \cup \{j\}$ we write $u_{i,j|A}$. With a slight abuse of notation, we denote the class of all elementary imsets by $\mathcal{E}$.

Primal feasibility in Theorem 3.5 requires that $\hat{\theta}$ satisfies (4.1), that is,

$$\langle \hat{\theta}, v \rangle \geq 0 \quad \text{for all } v \in \mathcal{E}. \tag{4.2}$$

The dual cone $\mathcal{C}^\vee$ is the cone in $\mathbb{R}^{\mathcal{X}}$ generated by all elementary imsets. Dual feasibility in Theorem 3.5 says that $\hat{\sigma}(x) > 0$ for all $x \in \mathcal{X}$ and

$$\hat{\sigma} - \bar{T} = \sum_{v \in \mathcal{E}} c_v v \quad \text{where } c_v \geq 0. \tag{4.3}$$

Although every element in $\mathcal{C}^\vee$ is a nonnegative combination of elementary imsets, such a combination is typically not unique. For example,

$$u_{1,2|3} + u_{1,3|\varnothing} = u_{1,3|2} + u_{1,2|\varnothing}.$$

In particular, the coefficients c_v above are not uniquely defined. But independent of the choice of these coefficients, the complementary slackness condition is equivalent to

$$\langle \hat{\theta}, \hat{\sigma} - \bar{T} \rangle = \sum_{v \in \mathcal{E}} c_v \langle \hat{\theta}, v \rangle = 0.$$

By (4.2), this holds if and only if

$$c_v \langle \hat{\theta}, v \rangle \quad = 0 \quad \text{for all } v \in \mathcal{E}. \tag{4.4}$$

We conclude that $\langle \hat{\theta}, v \rangle = 0$ for every v that appears in a nonnegative linear combination of the form (4.3). Therefore, we obtain the following result.

PROPOSITION 4.2. *Each equality in* (4.4) *corresponds to a context specific conditional independence statement where two variables are independent conditioned on a particular value of the remaining variables, as represented by an elementary imset.*

PROOF. Each inequality for a given elementary imset in (4.1) can be interpreted as a sign condition on a specific conditional correlation

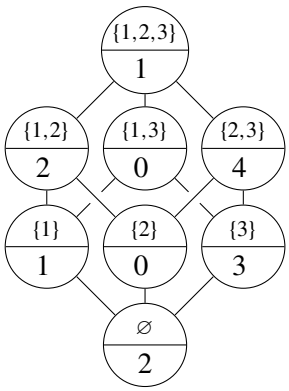
$$\operatorname{cov}(X_i, X_j \mid X_{V\setminus\{i,j\}} = x) \geq 0,$$

corresponding to an elementary imset. □

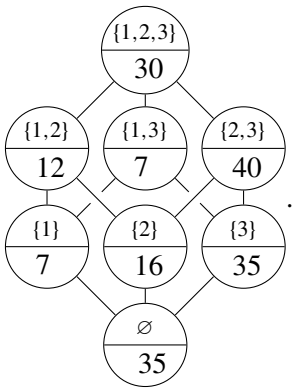
Note that when $d = 3$ there are six such constraints and these play an important role in the boundary decomposition of the latent class model [3]. To see how they appear in the description of a general binary latent class model, see [4]. In the following example, we show how this characterization of complementary slackness can be used to compute the MLE.

EXAMPLE 4.3. Let $d = 3$ and consider the sample represented by the diagram to the left in the following figure, where we again made use of the bijection between $\{-1, 1\}^3$ and the set of all subsets of $\{1, 2, 3\}$.

$$\bar{T} = \frac{1}{13} \cdot$$



$$\hat{\sigma} = \frac{1}{182} \cdot$$



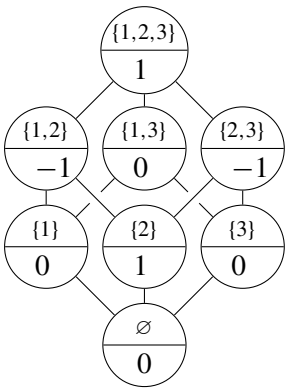
.

We claim that $\hat{\sigma}$ represented by the diagram on the right corresponds to the MLE. First, we check that $\hat{\sigma}$ is indeed MTP_2 by checking that $\hat{\sigma}(x \vee y)\hat{\sigma}(x \wedge x) - \hat{\sigma}(x)\hat{\sigma}(y) \geq 0$ for all six elementary pairs x, y. Up to the normalizing constant 182, these are

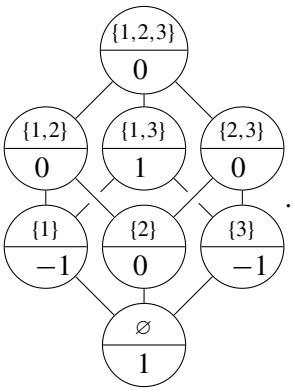
$$\begin{array}{ll}
\{1\},\{2\}: \ 12 \cdot 35 - 7 \cdot 16 > 0 & \{1,3\},\{2,3\}: \ 30 \cdot 35 - 7 \cdot 40 > 0 \\
\{1\},\{3\}: \ \mathbf{7 \cdot 35 - 7 \cdot 35 = 0} & \{1,2\},\{2,3\}: \mathbf{30 \cdot 16 - 12 \cdot 40 = 0}. \\
\{2\},\{3\}: 40 \cdot 35 - 16 \cdot 35 > 0 & \{1,2\},\{1,3\}: \ 30 \cdot 7 - 12 \cdot 7 > 0
\end{array}$$

This proves primal feasibility in Theorem 3.5. Dual feasibility is verified by the following diagram.

$$\hat{\sigma} - \bar{T} = \frac{16}{182} \cdot$$



$$+ \frac{7}{182} \cdot$$



.

In other words, $\hat{\sigma} - \bar{T} = \frac{16}{182} \cdot u_{1,3|2} + \frac{7}{182} \cdot u_{1,3|\varnothing} \in \mathcal{C}^\vee$. Complementary slackness follows by direct calculations. Note that the two nonzero generators in the decomposition of $\hat{\sigma} - \bar{T}$ correspond precisely to the MTP_2 inequalities for $\hat{\sigma}$ that hold as equalities. These equalities correspond to the conditional independence statement $1 \perp\!\!\!\perp 3 \mid 2$.

4.3. *Existence of the MLE.* In this section, we shall discuss problems associated with existence of the MLE for binary MTP_2 distributions, the main result being Theorem 4.5 which gives a simple necessary and sufficient condition for existence.

4.3.1. *Existence in the extended family.* To derive simple conditions for existence of the MLE within the exponential family of strictly positive binary distributions that are MTP_2, we consider estimation in the extended family where the strict positivity condition is relaxed and existence therefore guaranteed.

Let $\mathbb{P}(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$ and $\mathcal{P}_2$ the set of all totally positive binary distributions, that is,

$$\mathcal{P}_2 = \{p \in \mathbb{P}(\mathcal{X}) \mid \forall x, y \in \mathcal{X} : p(x \vee y)p(x \wedge y) \geq p(x)p(y)\}.$$

We note that $\mathcal{P}_2$ is compact and *geometrically convex*, that is,

$$p_1, p_2 \in \mathcal{P}_2 \implies c^{-1}\sqrt{p_1 p_2} \in \mathcal{P}_2$$

where

$$c := \sum_{x \in \mathcal{X}} \sqrt{p_1(x)p_2(x)} \leq 1$$

and $c < 1$ unless $p_1 = p_2$ by the Cauchy–Schwarz inequality.

For a lattice L, we say that a subset L' of L forms a *sublattice* of L if for any two $x, y \in L'$ it holds that $x \wedge y \in L'$ and $x \vee y \in L'$. Note that for any $p \in \mathcal{P}_2$ its support $\text{supp}(p) = \{x : p(x) > 0\}$ is always a sublattice of $\mathcal{X}$, since

$$p(x) > 0, p(y) > 0 \implies p(x \vee y)p(x \wedge y) \geq p(x)p(y) > 0.$$

Consider a sample $U = \{x^1, \ldots, x^n\}$ with likelihood function

$$L(p) = \prod_{i=1}^{n} p(x^i)$$

and let $\mathcal{L}(U)$ be the the smallest sublattice of $\mathcal{X}$ containing the sample U. We now show that the support of the MLE is given by $\mathcal{L}(U)$.

THEOREM 4.4. *The likelihood function attains its maximum over $\mathcal{P}_2$ in a unique point $\hat{p}$. Furthermore, it holds that* $\text{supp}(\hat{p}) = \mathcal{L}(U)$.

PROOF. Continuity of the likelihood function together with compactness of $\mathcal{P}_2$ ensures that the maximum is attained. To prove uniqueness, suppose for contradiction that $\hat{p}_1 \neq \hat{p}_2$ both maximize L. Then

$$L\big(c^{-1}\sqrt{\hat{p}_1 \hat{p}_2}\big) = c^{-n}\sqrt{L(\hat{p}_1)L(\hat{p}_2)} > L(\hat{p}_i)$$

contradicting that $\hat{p}_i$ were maximizers.

Finally, note that $U \subseteq \text{supp}(\hat{p})$ and hence $\mathcal{L}(U) \subseteq \text{supp}(\hat{p})$. We show $\mathcal{L}(U) \supseteq \text{supp}(\hat{p})$ by contradiction. Suppose $\mathcal{L}(U) \subsetneq \text{supp}(\hat{p})$, then we can construct $\tilde{p} \in \mathcal{P}_2$ such that $L(\tilde{p}) > L(\hat{p})$, which contradicts the fact that $\hat{p}$ is the MLE; namely, let $\tilde{p}$ be $\hat{p}$ projected onto $\mathcal{L}(U)$ and rescaled to be a probability mass function, that is, $\tilde{p}(x) \propto p(x)\mathbf{1}_{\mathcal{L}(U)}$. Then $\tilde{p} \in \mathcal{P}_2$ and $L(\tilde{p}) > L(\hat{p})$, which concludes the proof. □

4.3.2. *Existence of MLE in the binary exponential family.* The MLE exists in the binary exponential family if and only if the estimator $\hat{p}$ in the extended family $\mathbb{P}(\mathcal{X})$ has full support. Thus as a consequence of Theorem 4.4 we obtain the following result, where $U_{ij} = \{x_{ij}^1, \ldots, x_{ij}^n\}$ denotes the marginal sample induced on the pair $ij, i \neq j$.

THEOREM 4.5. *The MLE exists within the space of totally positive canonical parameters $\mathcal{K}_2$ (cf. Definition 3.1) if and only if $\mathcal{L}(U) = \mathcal{X}$. Furthermore, $\mathcal{L}(U) = \mathcal{X}$ if and only if every pair-marginal sample U_{ij} for $i, j \in V = \{1, \ldots, d\}$ has both of $(1, -1)$ and $(-1, 1)$ represented.*

PROOF. As mentioned, the MLE exists in the binary exponential family if and only if the estimator $\hat{p}$ in the extended family $\mathbb{P}(\mathcal{X})$ has full support. Thus, as a consequence of Theorem 4.4, the MTP_2 MLE exists if and only $\mathcal{L}(U) = \mathcal{X}$.

For the second statement we first prove the backward direction using the identification between $\mathcal{X}$ and subsets of V. Suppose every pair-marginal U_{ij} for $i, j \in V$ has both of $(1, -1)$ and $(-1, 1)$ represented. This means that for every i there is a set $x_{ij} \in U$ with $i \in x_{ij}$ and $j \notin x_{ij}$. But then

$$\{i\} = \bigcap_{j \in V \setminus i} x_{ij} \in \mathcal{L}(U) \quad \text{for all } i.$$

Since the set of all singletons $\{i\}$ for $i \in V$ generates the full lattice $\mathcal{X}$, we obtain $\mathcal{L}(U) = \mathcal{X}$ as desired.

We prove the forward direction by proving its contrapositive. Suppose there is a pair ij such that all sets $x \in U$ have the property that

$$i \in x \implies j \in x. \tag{4.5}$$

The set of subsets y satisfying (4.5) form a proper sublattice $\mathcal{L}' \subset \mathcal{X}$. Since $\mathcal{L}(U) \subseteq \mathcal{L}'$ we obtain that $\mathcal{L}(U) \neq \mathcal{X}$, which completes the proof. □

Theorem 1 in [33] states that the MLE in the MTP_2 Gaussian distribution exists if and only if all sample correlations are strictly less than one. Theorem 4.5 yields the analogous result for binary distributions. Indeed we have the following.

COROLLARY 4.6. *If the MLE exists within $\mathcal{K}_2$, then the empirical covariance matrix satisfies $S_{ij} < \sqrt{S_{ii} S_{jj}}$ for all $i \neq j$.*

PROOF. The empirical correlation matrix R has $|R_{ij}| = 1$ if and only if it holds for all $x \in U$ in the sample that $x_j = a x_i + b$. If both configurations $(1, -1)$ and $(-1, 1)$ are represented in U, this would imply $b - a = 1$ and $b + a = -1$ whereby $b = 0$, $a = -1$ and thus $R_{ij} = -1$ implying $S_{ij} < \sqrt{S_{ii} S_{jj}}$. □

Note that the converse is not true. If for two variables the sample is $U = \{(-1, -1), (1, -1), (1, 1)\}$, then the MLE does not exist according to Theorem 4.5, but we have $S_{11} = S_{22} = 8/9$ and $S_{12} = 4/9$; so the empirical correlation is equal to $1/2$.

As another example, consider the case $d = 3$. Then the vectors $(1, -1, -1)$, $(-1, 1, -1)$, $(-1, -1, 1)$ generate all of $\{-1, 1\}^3$ and hence every sample supported on these three points will admit a unique MLE under the MTP_2 constraint. This set is minimal in the sense that it cannot be reduced; none of its subsets generates $\mathcal{X}$. There are also minimal generating subsets of size four, for example, $(1, 1, -1)$, $(1, -1, -1)$, $(-1, -1, 1)$, $(-1, 1, 1)$. For general d, a minimal generating set of $\{-1, 1\}^d$ is of order $\mathcal{O}(d)$ and there always exists a minimal

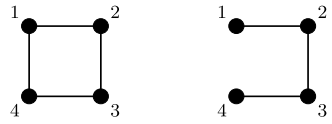


FIG. 1. *A cycle (left) and a chain (right) with four vertices.*

generating set of size exactly d. Hence for binary MTP_2 distributions d samples can be sufficient for existence of the MLE. This is in sharp contrast with unrestricted binary exponential families, where the MLE exists only if *all* 2^d states are observed at least once.

While the MLE in Example 4.3 could be computed by hand, calculations get intractable rather quickly. The following example is sufficiently complicated that it cannot easily be calculated by hand, but still simple enough so that numerical optimization using the algorithm developed in [9] yields the provably exact optimum.

EXAMPLE 4.7. Moussouris [29] provided a now classical example of a distribution q that is globally Markov to its dependence graph but does not factorize; cf. [26], Example 3.10. The distribution in this example is uniformly supported on eight points

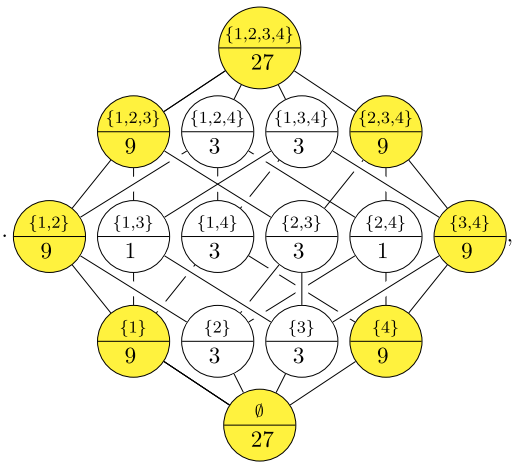
$$\begin{array}{llll}(-1,-1,-1,-1) & (1,-1,-1,-1) & (1,1,-1,-1) & (1,1,1,-1)\\ (-1,-1,-1,1) & (-1,-1,1,1) & (-1,1,1,1) & (1,1,1,1).\end{array}$$

This distribution is globally Markov with respect to the 4-cycle in Figure 1 (left), and we shall consider these eight points as constituting a sample of size eight. The MLE for this graphical model as an exponential family does not exist. Note that the sample distribution is not MTP_2, since, for example, the inequality

$$p(1,-1,-1,1)p(-1,-1,-1,-1) \geq p(1,-1,-1,-1)p(-1,1,-1,-1)$$

does *not* hold. On the other hand, since the conditions of Theorem 4.5 are satisfied, the MLE $\hat{\sigma}$ under MTP_2 exists. It is represented by the following diagram, where the highlighted nodes correspond to the eight points supported by the sample.

$\hat{\sigma} = \frac{1}{128} \cdot$



Primal feasibility of $\hat{\sigma}$ is verified by the following inequalities, one for each of the 24 elementary pairs (labeled by sets $\{i\} \cup A$ and $\{j\} \cup A$). Up to the normalizing constant 128, these

are:

$$\begin{array}{llll}
\{1\},\{2\}: & 9\cdot 27-9\cdot 3>0 & \{1,3\},\{2,3\}: & 9\cdot 3-1\cdot 3>0\\
\{1,4\},\{2,4\}: & 3\cdot 9-3\cdot 1>0 & \{1,3,4\},\{2,3,4\}: & 27\cdot 9-3\cdot 9>0\\
\{1\},\{3\}: & \mathbf{1\cdot 27-9\cdot 3=0} & \{1,2\},\{2,3\}: & \mathbf{9\cdot 3-9\cdot 3=0}\\
\{1,4\},\{3,4\}: & \mathbf{3\cdot 9-3\cdot 9=0} & \{1,2,4\},\{2,3,4\}: & \mathbf{27\cdot 1-3\cdot 9=0}\\
\{1\},\{4\}: & \mathbf{3\cdot 27-9\cdot 9=0} & \{1,2\},\{2,4\}: & \mathbf{3\cdot 3-9\cdot 1=0}\\
\{1,3\},\{3,4\}: & \mathbf{3\cdot 3-1\cdot 9=0} & \{1,2,3\},\{2,3,4\}: & \mathbf{27\cdot 3-9\cdot 9=0}\\
\{2\},\{3\}: & 3\cdot 27-3\cdot 3>0 & \{1,2\},\{1,3\}: & 9\cdot 9-9\cdot 1>0\\
\{2,4\},\{3,4\}: & 9\cdot 9-1\cdot 9>0 & \{1,2,4\},\{1,3,4\}: & 27\cdot 3-3\cdot 3>0\\
\{2\},\{4\}: & \mathbf{1\cdot 27-3\cdot 9=0} & \{1,2\},\{1,4\}: & \mathbf{3\cdot 9-9\cdot 3=0}\\
\{2,3\},\{3,4\}: & \mathbf{9\cdot 3-3\cdot 9=0} & \{1,2,3\},\{1,3,4\}: & \mathbf{27\cdot 1-9\cdot 3=0}\\
\{3\},\{4\}: & 9\cdot 27-3\cdot 9>0 & \{1,3\},\{1,4\}: & 3\cdot 9-1\cdot 3>0\\
\{2,3\},\{2,4\}: & 9\cdot 3-3\cdot 1>0 & \{1,2,3\},\{1,2,4\}: & 27\cdot 9-9\cdot 3>0
\end{array}$$

Quite surprisingly, the MLE is therefore still globally Markov to the 4-cycle even though these constraints were not explicitly enforced. Moreover, $\hat{\sigma}$ satisfies an additional conditional independence relation, namely $1 \perp\!\!\!\perp 4|\{2,3\}$, and so it is Markov to the smaller graph in Figure 1 (right).

There are many equivalent ways to write the vector $\hat{\sigma}-\bar{T}$. The most canonical is the one using all twelve elementary imsets allowed by the complementary slackness condition (4.4), that is, the ones corresponding to boldfaced rows above:

$$\begin{aligned}
\hat{\sigma}-\bar{T} = {} & \frac{3}{128}\cdot u_{1,3|\varnothing}+\frac{1}{128}\cdot u_{1,3|2}+\frac{1}{128}\cdot u_{1,3|4}+\frac{3}{128}\cdot u_{1,3|2,4}\\
& +\frac{3}{128}\cdot u_{2,4|\varnothing}+\frac{1}{128}\cdot u_{2,4|1}+\frac{1}{128}\cdot u_{2,4|3}+\frac{3}{128}\cdot u_{2,4|1,3}\\
& +\frac{5}{128}\cdot u_{1,4|\varnothing}+\frac{5}{128}\cdot u_{1,4|2}+\frac{5}{128}\cdot u_{1,4|3}+\frac{5}{128}\cdot u_{1,4|2,3}.
\end{aligned}$$

Each of the vectors $u_{i,j|A}$ above is a generator of $\mathcal{C}^\vee$ and so $\hat{\sigma}-\bar{T}\in\mathcal{C}^\vee$.

REMARK 4.8. To show that $\hat{\sigma}-\bar{T}$ lies in $\mathcal{C}^\vee$, it is enough to express it as a nonnegative combination of vectors $T(x\wedge y)+T(x\vee y)-T(x)-T(y)$ for *arbitrary* pairs $x,y\in\mathcal{X}$. This follows directly from [34], Proposition 4.2.

Note that in the above examples the MLEs correspond to models satisfying conditional independence statements. However, in general the MLE will satisfy a set of context specific conditional independence statements that may not lead to full conditional independences. In the following subsection, we consider binary MTP_2 models that satisfy conditional independence relations given by a graphical model.

4.4. *Totally positive graphical models for binary variables.* Given a graph $G=(V,E)$, let $\mathcal{P}_2(G)$ denote the set of distributions in $\mathcal{P}_2$ that lie in the completion of the exponential family ([7], pp. 154–155) for the graphical model over G, that is,

$$\mathcal{P}_2(G)=\mathcal{P}_2\cap M_E(G),$$

where $M_E(G)$ denotes the set of *extended Markov* distributions ([26], p. 40) obtained as limits of factorizing distributions; see also [20]. We note that $\mathcal{P}_2(G)$ is compact and geometrically convex (see, e.g., [26], p. 73); hence the MLE over $\mathcal{P}_2(G)$ exists and is unique.

We first need a lemma to identify when binary MTP_2 distributions $p\in\mathcal{P}_2$ have full support based on their marginals. These results are critical for this section in order to identify when the MLE of a binary distribution that is Markov over a graph has full support.

LEMMA 4.9. *Let $p \in \mathcal{P}_2$ and let $x \in \mathcal{X}$. Suppose $p_{ij}(x_i, x_j) > 0$ for all pairs i, j then $p(x) > 0$.*

PROOF. For every i, j let $y^{(ij)} \in \text{supp}(p)$ such that $y_{ij}^{(ij)} = (x_i, x_j)$. Let A/B be the partition of V such that $x_i = -1$ for $i \in A$ and $x_i = 1$ on B. For each $i \in A$, define $z^{(i)} = \max_{j \in B} y^{(ij)}$. By construction, $z_i^{(i)} = -1$ and $z_B^{(i)} = (1, \ldots, 1)$. Moreover, $z^{(i)} \in \text{supp}(p)$ because $\text{supp}(p)$ is a lattice. Since $x = \min_{i \in A} z^{(i)}$, $x \in \text{supp}(p)$ again because the support of p is a lattice. □

COROLLARY 4.10. *If $p \in \mathcal{P}_2$, then p has full support $\mathcal{X}$ if and only if each pair-margin p_{ij} has full support.*

The following result extends Theorem 4.5 to binary graphical models and relaxes the pair-marginal condition to be necessary only for pairs of neighbours in the graph G. As before $U_{ij} = \{x_{ij}^1, \ldots, x_{ij}^n\}$ denotes the pair-marginal sample for the pair ij.

THEOREM 4.11. *If every pair marginal sample U_{ij} along edges $ij \in E$ has both of $(1, -1)$ and $(-1, 1)$ represented, then the unique MLE $\hat{p} \in \mathcal{P}_2(G)$ has full support.*

The proof makes use of the fact that the support of $\hat{p}$, denoted by $\text{supp}(\hat{p})$, is a lattice since $\hat{p} \in \mathcal{P}_2$. In addition, since $\hat{p} \in M_E(G)$, $\hat{p}$ also satisfies the global, local and pairwise Markov properties w.r.t. G ([26], p. 42, (3.16). In particular, the proof relies on the following two lemmas.

LEMMA 4.12. *If the pair marginal sample U_{ij} has both of $(1, -1)$ and $(-1, 1)$ represented for all $ij \in E$, then $\text{supp}(\hat{p}_{ij}) = \{-1, 1\}^2$ for all $ij \in E$.*

PROOF. The MTP_2 property is closed under taking marginals (see [22]). So if $\hat{p}$ is MTP_2, so are its marginals $\hat{p}_{ij}$. Thus $\text{supp}(\hat{p}_{ij})$ is a lattice containing U_{ij}. As a consequence, if U_{ij} has both of $(1, -1)$ and $(-1, 1)$ represented, then $\text{supp}(\hat{p}_{ij}) = \{-1, 1\}^2$, which completes the proof. □

Denoting by ∂i the neighbors of node $i \in V$ in G, the following lemma will be needed for showing that $\text{supp}(\hat{p}_{ij}) = \{-1, 1\}^2$ for *all* pairs ij and not only the pairs $ij \in E$.

LEMMA 4.13. *Suppose that every pair marginal sample U_{ij} along edges $ij \in E$ has both of $(1, -1)$ and $(-1, 1)$ represented. If $\hat{p}_{\partial i}(x_{\partial i}) > 0$ for some $x_{\partial i}$, then $\hat{p}_{i \cup \partial i}(x_{i \cup \partial i}) > 0$ for every x_i.*

PROOF. Since $\hat{p}_{\partial i}(x_{\partial i}) > 0$, clearly $\hat{p}_{i \cup \partial i}(x_{i \cup \partial i}) > 0$ for some x_i, say $x_i = 1$. We need to show that $\hat{p}_{i \cup \partial i}(y_{i \cup \partial i}) > 0$ also if $y_i = -1$ and $y_{\partial i} = x_{\partial i}$. Let $z_{i \cup \partial i}$ be such that $z_i = -1$ and $z_{\partial i} = (1, \ldots, 1)$. Since $\hat{p} \in \mathcal{P}_2(G)$, its support is a lattice and the same applies to each margin of $\hat{p}$. Because

$$y_{i \cup \partial i} = x_{i \cup \partial i} \wedge z_{i \cup \partial i},$$

to show that $y_{i \cup \partial i}$ lies in the support of $\hat{p}_{i \cup \partial i}$ it is sufficient to show that this holds for $z_{i \cup \partial i}$. By the assertion, for each $j \in \partial i$ the edge-margin U_{ij} has $(-1, 1)$ represented. In particular, there is a point $u^{(j)} \in \mathcal{X}$ such that $u_i^{(j)} = -1$ and $u_j^{(j)} = 1$. The support of $\hat{p}$ necessarily contains all elements in U and hence $\hat{p}(u^{(j)}) > 0$ for all $j \in \partial i$. Let u be the elementwise

maximum of all $u^{(j)}$. This point lies in $\operatorname{supp}(\hat{p})$ because it forms a lattice. By construction, $u_{i\cup\partial i} = z_{i\cup\partial i}$, which proves that $z_{i\cup\partial i}$ (and hence also $y_{i\cup\partial i}$) lies in the support of $\hat{p}_{i\cup\partial i}$. The proof for the case where $x_i = -1$ is analogous. □

We are now ready to provide the proof of Theorem 4.11.

PROOF OF THEOREM 4.11. From Corollary 4.10, it follows that $\hat{p}$ has full support if and only if the marginal support $\operatorname{supp}(\hat{p}_{ij})$ is full for all $i, j \in V$. When $ij \in E$, this follows from Lemma 4.12. Next, consider a pair $ij \notin E$. Since $\hat{p} \in M_E(G)$, it satisfies the local Markov property with respect to G. Hence for any $x_i, x_j \in \{-1, 1\}$, it holds that

$$\begin{aligned}\hat{p}_{ij}(x_i, x_j) &= \sum_{x_{\partial i\cup\partial j}} \hat{p}(x_i, x_j \mid x_{\partial i\cup\partial j})\hat{p}(x_{\partial i\cup\partial j}) \\ &= \sum_{x_{\partial i\cup\partial j}} \hat{p}(x_i \mid x_{\partial i})\hat{p}(x_j \mid x_{\partial j})\hat{p}(x_{\partial i\cup\partial j}).\end{aligned}$$

Since there is at least one $x_{\partial i\cup\partial j}$ in the support of $\hat{p}_{\partial i\cup\partial j}$, then by Lemma 4.13 both of $\hat{p}(x_i, x_{\partial i})$ and $\hat{p}(x_j, x_{\partial j})$ are strictly positive and hence also the corresponding summand. It follows that $\hat{p}_{ij}(x_i, x_j) > 0$, as desired. □

Theorem 4.11 provides conditions for the existence of the MLE in the underlying exponential family, which we denote by $\mathcal{K}_2(G)$, consisting of all points in $\mathcal{P}_2(G)$ with full support.

COROLLARY 4.14. *If G is bipartite, then the minimal sample size required for existence of the MLE is $n = 2$. More generally, for arbitrary graphs the minimal sample size for existence of the MLE is of the order of the maximal clique size.*

Hence the minimal sample size for existence of the MLE goes from 2^d for unrestricted binary distributions, to d for MTP_2 binary distributions, to $\mathcal{O}$(maximal clique size) for MTP_2 binary distributions on graphs, including Ising models. In the following subsection, we consider a special class of binary distributions that contain as prominent examples Ising models without external field and show that the minimal sample size for existence of the MLE can be further reduced.

4.5. *Symmetric binary distributions.* A distribution p over $\mathcal{X} = \{-1, 1\}^d$ is *symmetric* (or *palindromic*) if $p(x) = p(-x)$ for all $x \in \mathcal{X}$. Distributions of this form have been studied, for example, in [28] and also appear in statistical physics in the context of spin models with no external field. If $X = (X_1, \ldots, X_d)$ has a symmetric distribution, then $\mathbb{E}X_i = 0$ and $\operatorname{var}(X_i) = 1$ for all $i = 1, \ldots, d$. As a consequence, the covariance matrix and the correlation matrix of X coincide. Note also that symmetry translates into linear constraints $\theta(x) = \theta(-x)$ for all $x \in \mathcal{X}$ on the canonical parameters of the binary exponential family. Hence symmetric distributions with full support form themselves an exponential family. In the following, we characterize existence of the MLE for symmetric binary distributions.

Let as before $U = \{x^1, \ldots, x^n\}$ denote a random sample. Let $\mathcal{A}(U)$ denote the smallest algebra generated by U, that is, the smallest subset of $\mathcal{X}$ that contains U and is closed under the lattice operations $\wedge, \vee$ and the complement $x \mapsto -x$. For a family of distributions $\mathcal{P}$, we let $\mathcal{P}^s$ denote the set of symmetric distributions in $\mathcal{P}$ and $U^s = U \cup -U$ be the symmetrized sample.

PROPOSITION 4.15. *If $\mathcal{P}$ is geometrically convex, then the MLE $\hat{p}_s$ under $\mathcal{P}^s$ based on a sample U exists in $\mathcal{P}^s$ if and only if the MLE $\tilde{p}_s$ under $\mathcal{P}$ based on the symmetrized sample U^s exists. In this case, it holds that $\hat{p}_s = \tilde{p}_s$.*

PROOF. Note that for any $p \in \mathcal{P}$, the likelihood function satisfies

$$L(p; U^s) = \prod_{x \in \mathcal{X}} p(x)^{n(x)+n(-x)} = L(\check{p}; U^s),$$

where $\check{p}(x) = p(-x)$, and $n(x) = |\{i \in 1, \ldots, n : x_i = x\}|$ are the empirical counts in the sample U. Since $\mathcal{P}$ is geometrically convex, a maximizer $\tilde{p}_s$ of $L(p; U^s)$ is unique; thus $\tilde{p}_s(x) = \tilde{p}_s(-x)$, and hence $\tilde{p} \in \mathcal{P}^s$. Note also that for any $p_s \in \mathcal{P}^s$ we have

$$L(p_s; U)^2 = L(p_s; U^s).$$

So any maximizer of $L(p_s; U)$ over $\mathcal{P}^s$ is also a maximizer of $L(p_s; U^s)$ and vice-versa. Finally, the uniqueness implies that $\hat{p}_s = \tilde{p}_s$, as desired. □

By combining Proposition 4.15 with Theorem 4.5, we obtain the following corollary on the existence of the MLE for symmetric binary distributions.

COROLLARY 4.16. *The MLE for a symmetric binary exponential family exists if and only if $\mathcal{A}(U) = \mathcal{X}$. Furthermore, $\mathcal{A}(U) = \mathcal{X}$ if and only if for every pair ij the event $\{X_i \neq X_j\}$ is represented in the sample.*

Finally, as a consequence we obtain the following corollary as an application of Theorem 4.11 to symmetric binary distributions on graphs defined as $\mathcal{P}_2^s(G) := \mathcal{P}_2(G) \cap \mathcal{P}_2^s$.

COROLLARY 4.17. *If the event $\{X_i \neq X_j\}$ is represented in every pair marginal sample U_{ij}, then the MLE $\hat{p}$ in the family $\mathcal{P}_2^s(G)$ has full support.*

REMARK 4.18. We note again the remark to Theorem 1 in [33] which states that the MLE in an MTP_2 Gaussian distribution exists if and only if all sample correlations are strictly less than one. Corollary 4.17 implies that *exactly the same is true for symmetric binary distributions.* Interestingly, while for (nontrivial, i.e., with at least one edge) Gaussian graphical models sample size equal to two is necessary and sufficient for existence of the MLE (with probability 1) [25], as a consequence of Corollary 4.14 and Corollary 4.17, the MLE for a symmetric binary distribution on a bipartite graph may have full support for sample size equal to one.

5. Totally positive Ising models. In this section, we study maximum likelihood estimation in *Ising models*, a special class of binary distributions that form a quadratic exponential family. An algorithm for calculating the MLE $\hat{p}$ for general binary MTP_2 distributions was developed in [9]. In Section 5.2, we develop an algorithm analogous to *iterative proportional scaling* (*IPS*) for the special case of Ising models under MTP_2. In addition, we discuss the special case of MTP_2 Ising models with no external field, which forms a symmetric exponential family. Such distributions can be seen as a proxy to Gaussian distributions and in Section 5.3 we discuss their similarities and differences.

Since Ising models form a quadratic exponential family, their probability mass function is of the form

$$p(x; h, J) = \exp\big(h^T x + x^T J x/2 - A(h, J)\big), \tag{5.1}$$

with $h \in \mathbb{R}^d$ and $J \in \mathbb{S}_0^d$, where $\mathbb{S}_0^d$ is the set of symmetric matrices in $\mathbb{R}^{d \times d}$ with $J_{ii} = 0$ for all i, ensuring minimality of the representation; see (3.3). We let $\mathcal{I}_2$ be the set of Ising models above that are also MTP_2, that is, where $J_{ij} \geq 0$ for all $i \neq j$.

Let $\theta = (h, J)$ denote the canonical parameters. We make the following two important observations regarding the canonical parameters. For any $i, j \in V$ let $A = V \setminus \{i, j\}$. Then the

corresponding conditional log-odds ratios are all equal; more precisely, denote by $x, y \in \mathcal{X}$ any two points satisfying $x_A = y_A$, $x_i = y_j = 1$, and $x_j = y_i = -1$, then

$$\log\left(\frac{p(x \vee y)p(x \wedge y)}{p(x)p(y)}\right) = 4J_{ij}. \tag{5.2}$$

This is another way of confirming that an Ising model defined by (h, J) is MTP_2 if and only if $J \in \mathbb{S}_+^d \cap \mathbb{S}_0^d$; see Proposition 3.6. In addition, note that for any x with $x_i = 1$ and y equal to x up to the i'th coordinate, then

$$\log\left(\frac{p(x)p(-y)}{p(-x)p(y)}\right) = 4h_i. \tag{5.3}$$

The sufficient statistics based on the observations $x^1, \ldots, x^n$ are the first- and second-order moments

$$(\bar{x}, M) := \frac{1}{n}\left(\sum_{i=1}^n x^i, \sum_{i=1}^n x^i (x^i)^T\right).$$

Strictly speaking, we should ignore the diagonal elements of M, but since they are all deterministically equal to 1, this does not matter for the following considerations. In addition, for a graphical Ising model on $G = (V, E)$—that is, where we assume $J_{ij} = 0$ unless $ij \in E$—the entries M_{ij} for $ij \notin E$ should be ignored. The associated *mean value parameters* are

$$(\mu, \Xi) := \left(\mathbb{E}_\theta X, \mathbb{E}_\theta X X^T\right).$$

5.1. *Existence of the MLE for totally positive Ising models.* Theorem 4.11 can be specialized to the quadratic case, that is, when also the Ising model is assumed. The condition for existence is here unchanged compared to the general Markov case. For an undirected graph $G = (V, E)$, let $\mathcal{I}_2(G)$ be the family of totally positive Ising models that are Markov w.r.t. G, that is, where $J_{ij} = 0$ unless $ij \in E$. We then have the following.

THEOREM 5.1. *If every pair marginal sample U_{ij} along edges $ij \in E$ has both of $(1, -1)$ and $(-1, 1)$ represented, then the MLE $\hat{p} \in \mathcal{I}_2(G)$ is unique and has full support.*

PROOF. By Theorem 4.11, the MLE exists within the convex exponential family $\mathcal{P}_2(G)$. Since $\mathcal{I}_2(G)$ is an exponential subfamily of that, the MLE also exists within $\mathcal{I}_2(G)$. □

In the following, we shall develop an algorithm for calculating the MLE in MTP_2 Ising model.

5.2. *IPS algorithm for computing the MLE.* The standard IPS algorithm (see [26], p. 82) for computing the MLE without the MTP_2 restriction works by cycling through all pairs $ij \in E$ and optimizing the likelihood function when fixing the values of all canonical parameters associated with variables other than the given pair, namely

$$h^{-ij} := (h_v, v \in V \setminus \{i, j\}), \qquad J^{-ij} := (J_{uv}, u, v \in V \setminus \{i, j\}).$$

Dually, this corresponds to fitting the mean value parameters associated with i, j to their empirically observed values, that is,

$$\mu_i = \bar{x}_i, \qquad \mu_j = \bar{x}_j, \qquad \Xi_{ij} = M_{ij}.$$

If the MLE exists, then this algorithm is known to converge to the MLE (see [26], p. 82). We next extend this algorithm to MTP_2 Ising models.

Let e_{ij} denote the empirical distribution of (X_i, X_j). Note that this distribution depends on the sufficient statistics through the formula

$$e_{ij}(1,1) = (1 + \bar{x}_i + \bar{x}_j + M_{ij})/4, \qquad e_{ij}(1,-1) = (1 + \bar{x}_i - \bar{x}_j - M_{ij})/4,$$
$$e_{ij}(-1,1) = (1 - \bar{x}_i + \bar{x}_j - M_{ij})/4, \qquad e_{ij}(-1,-1) = (1 - \bar{x}_i - \bar{x}_j + M_{ij})/4.$$

We now assume that $e_{ij}(1,-1) > 0$ and $e_{ij}(-1,1) > 0$ for all $ij \in E$, which ensures that $-1 < \bar{x}_i < 1$ for all $i \in V$ and that the MLE has full support; see Theorem 4.11. By Corollary 3.7 and the following paragraph, for edges where $S_{ij} = M_{ij} - \bar{x}_i\bar{x}_j < 0$, it holds that $J_{ij} = 0$. For the other edges, it holds that

$$e_{ij}(1,1) \geq (1 + \bar{x}_i + \bar{x}_j + \bar{x}_i\bar{x}_j)/4 = (1 + \bar{x}_i)(1 + \bar{x}_j)/4 > 0$$

and, similarly, $e_{ij}(-1,-1) \geq (1 - \bar{x}_i)(1 - \bar{x}_j) > 0$.

The IPS algorithm is initialized in any point inside the model such as the uniform distribution or the distribution where all variables are mutually independent with mean $\hat{\mu} = \bar{x}$. The update for the edge $ij \in E$ can be expressed as

$$p(x) \leftarrow p(x)\frac{e_{ij}(x_i, x_j)}{p_{ij}(x_i, x_j)} = p(x_{-ij} \mid x_i, x_j)e_{ij}(x_i, x_j) = p(x)q_{ij}(x_i, x_j). \tag{5.4}$$

Using (5.2), we easily verify that J_{ij} is the only entry of J affected by this update. Exploiting that $q_{ij}(x_i, x_j) > 0$, we can define

$$\Delta_{ij} := \frac{1}{4}\log\frac{q_{ij}(1,1)q_{ij}(-1,-1)}{q_{ij}(1,-1)q_{ij}(-1,1)}. \tag{5.5}$$

Using a mixed parametrization (see [7]) with (μ_i, μ_j, J_{ij}) and canonical para- meters for all other indices, the update step can equivalently be expressed as

$$J_{ij} \leftarrow J_{ij} + \Delta_{ij}, \quad \mu_i \leftarrow \bar{x}_i, \quad \mu_j \leftarrow \bar{x}_j,$$

where all other entries of (h, J) remain unchanged.

To ensure the MTP_2 constraint, it is natural to replace J_{ij} with zero if the update becomes negative and then recalculate (h_i, h_j) to comply with the requirement $(\mu_i, \mu_j) = (\bar{x}_i, \bar{x}_j)$.

Alternatively, we can express the update in terms of mean value parameters by letting $\hat{\Xi}_{ij} \leftarrow M_{ij} + \lambda^*$. To compute λ^*, define $e^*_{ij} = e_{ij}(\lambda^*)$ by

$$e^*_{ij}(1,1) = (1 + \bar{x}_i + \bar{x}_j + \hat{\Xi}_{ij})/4 = e_{ij}(1,1) + \lambda^*/4,$$
$$e^*_{ij}(1,-1) = (1 + \bar{x}_i - \bar{x}_j - \hat{\Xi}_{ij})/4 = e_{ij}(1,-1) - \lambda^*/4,$$
$$e^*_{ij}(-1,1) = (1 - \bar{x}_i + \bar{x}_j - \hat{\Xi}_{ij})/4 = e_{ij}(-1,1) - \lambda^*/4,$$
$$e^*_{ij}(-1,-1) = (1 - \bar{x}_i - \bar{x}_j + \hat{\Xi}_{ij})/4 = e_{ij}(-1,-1) + \lambda^*/4,$$

and define $q^*_{ij} = e^*_{ij}/p_{ij}$. Then λ^* is given by the solution to the equation

$$\Delta_{ij}(\lambda) = -J_{ij}, \tag{5.6}$$

where

$$\Delta_{ij}(\lambda^*) = \frac{1}{4}\log\frac{q^*_{ij}(1,1)q^*_{ij}(-1,-1)}{q^*_{ij}(1,-1)q^*_{ij}(-1,1)}.$$

Note that $\Delta_{ij}(\lambda)$ is strictly increasing in λ, $\Delta_{ij}(0) < -J_{ij}$, and $\Delta_{ij}(\lambda) \to \infty$ for $\lambda \to \min(e_{ij}(1,-1), e_{ij}(-1,1))$. Hence there is a unique solution λ^* with $\lambda^* > 0$. Letting $x = \lambda/4$, then (5.6) becomes

$$\log\left(\frac{(e_{ij}(1,1) + x)(e_{ij}(-1,-1) + x)}{(e_{ij}(-1,1) - x)(e_{ij}(1,-1) - x)}\right) = -\log\left(\frac{p_{ij}(1,1)p_{ij}(-1,-1)}{p_{ij}(-1,1)p_{ij}(1,-1)}\right) - 4J_{ij},$$

Algorithm 1 IPS-type algorithm for computing the MLE in MTP_2 Ising models

input: Sample moments $(\bar{x}, M)$, a graph $G = (V, E)$, and precision ϵ.
output: The MLE $(\hat{p}, \hat{G}, \hat{\mu}, \hat{\Xi})$.

initialize $\mu = \bar{x}$; $p(x) = 2^{-|V|} \prod_{v \in V}(1 + x_v \mu_v)$ for all $x \in \mathcal{X}$; $\Xi = \mathbf{I}$;
initialize $E^+ = \{uv \in E \mid M_{uv} > \bar{x}_u \bar{x}_v\}$; $\hat{E} = \varnothing$;

repeat
  for $ij \in E^+$ **do**
    calculate Δ_{ij} by (5.5);
    calculate J_{ij} by (5.2);
    if $\Delta_{ij} + J_{ij} > 0$ **then**
      update p by (5.4);
      $\hat{E} \leftarrow \hat{E} \cup \{ij\}$;
    else
      solve $\Delta_{ij}(\lambda) = -J_{ij}$;
      update p by (5.7);
      $\hat{E} \leftarrow \hat{E} \setminus \{ij\}$;
    end if
  end for
  calculate (μ, Ξ) from p;
until $\max_{v \in V} |\hat{\mu}_v - \bar{x}_v| < \epsilon$ **and** $\min_{uv \in E}(\Xi_{uv} - M_{uv}) \geq 0$ **and** $\max_{uv \in \hat{E}} |\Xi_{uv} - M_{uv}| < \epsilon$;
return $p, \hat{G} = (V, \hat{E}), \mu, \Xi$.

or equivalently,

$$\frac{(e_{ij}(1,1)+x)(e_{ij}(-1,-1)+x)}{(e_{ij}(-1,1)-x)(e_{ij}(1,-1)-x)} = \frac{p_{ij}(1,1)p_{ij}(-1,-1)}{p_{ij}(-1,1)p_{ij}(1,-1)} \cdot e^{-4J_{ij}}.$$

Denoting the right-hand side of the above equation by R, multiplying both sides by $(e_{ij}(-1,1)-x)(e_{ij}(1,-1)-x)$, and moving everything to the left, we obtain a quadratic equation $ax^2 + bx + c = 0$ with $a = 1 - R$,

$$b = e_{ij}(1,1) + e_{ij}(-1,-1) + R\big(e_{ij}(-1,1) + e_{ij}(1,-1)\big),$$
$$c = e_{ij}(1,1)e_{ij}(-1,-1) - R\big(e_{ij}(-1,1)e_{ij}(1,-1)\big).$$

Hence the solution $\lambda^* = 4x^*$ is given by taking x^* to be the positive root of this quadratic equation. Using λ^*, we can then update $p(x)$ as follows:

$$p(x) \leftarrow p(x)\frac{e^*_{ij}(x_i, x_j)}{p_{ij}(x_i, x_j)}. \tag{5.7}$$

The full procedure is presented in Algorithm 1. In the following theorem, we show that this procedure indeed converges to the MLE (if it exists).

THEOREM 5.2. *If the MLE for the* MTP_2 *Ising model on the undirected graph* $G = (V, E)$ *exists, then the output of Algorithm 1 converges to the MLE for* $\epsilon \to 0$.

PROOF. Let (h, J) denote the canonical parameters of the exponential family. Then the log-likelihood function satisfies

$$-\frac{1}{n}\log L(h, J) = \log c(h, J) - h^T \bar{x} - \text{tr}(JM)/2,$$

where $c(h, J)$ is the normalizing constant of the exponential family. We fix a value (h^0, J^0) with $J^0_{uv} \geq 0$ and consider the following restricted convex optimization problem:

$$\begin{aligned} \underset{(h,J)}{\text{minimize}} \quad & \log c(h, J) - h^T \bar{x} - \operatorname{tr}(JM)/2 \\ \text{subject to} \quad & J_{ij} \geq 0, h_u = h^0_u, u \in V \setminus \{i, j\}, J_{uv} = J^0_{uv} \text{ for } uv \neq ij. \end{aligned}$$

Exploiting that most entries of (h, J) are fixed, this problem is equivalent to

$$\begin{aligned} \underset{(h_i,h_j,J_{ij})}{\text{minimize}} \quad & \log c(h, J) - h_i \bar{x}_i - h_j \bar{x}_j - J_{ij} M_{ij} \\ \text{subject to} \quad & J_{ij} \geq 0, \end{aligned}$$

where the fixed values $h_u = h^0_u$, $u \in V \setminus \{i, j\}$ and $J_{uv} = J^0_{uv}$ for $uv \neq ij$ enter into the function $\log c(h, J)$. Since also this subfamily is a convex exponential family, the solution to this optimization problem is uniquely determined by:

(i) Primal feasibility: $\hat{J}_{ij} \geq 0$
(ii) Dual feasibility: $\hat{\mu}_i = \bar{x}_i$, $\hat{\mu}_j = \bar{x}_j$, and $\hat{\Xi}_{ij} \geq M_{ij}$,
(iii) Complementary slackness: $(\hat{\Xi}_{ij} - M_{ij})\hat{J}_{ij} = 0$.

Thus, if $J^0_{ij} + \Delta_{ij} \geq 0$ we update as in (5.4). Else we update as in (5.7).

Note that every step of the algorithm maximizes the likelihood over a section. In addition, any fixed point of the algorithm satisfies the conditions in Corollary 3.7, and hence must be equal to the unique MLE. Furthermore, the updates depend continuously on p. Hence the algorithm is an instance of iterative partial maximization as described in [26], p. 230, and is therefore convergent with the unique MLE as limit. □

We note a computational issue with Algorithm 1. As stated above, the algorithm requires visiting all possible states $x \in \mathcal{X}$, which becomes computationally prohibitive for large d as the computational effort is then exponential in d. This problem can be overcome by an appropriate use of probability propagation as in [21]. More precisely, instead of representing p by its *values* $p(x), x \in \mathcal{X}$, we represent p by a set of *potentials* $\psi_{ij}, ij \in E$, such that

$$p(x) \propto \prod_{ij \in E} \psi_{ij}(x_i, x_j) = \prod_{ij \in E} \exp(x_i x_j J_{ij}).$$

Whenever a marginal $p(x_u, x_v)$ is required for the update, it is calculated from J by probability propagation as, for example, described in [15]. Then instead of updating p itself, the update (5.4) or (5.7) is performed by updating J only. This reduces the computational effort to become linear in the maximal clique size of G rather than d.

Finally, note that since the algorithm runs entirely in terms of probabilities $p(x)$, a simple modification of the algorithm as in [24, 26] guarantees convergence even when the MLE does not exist within the exponential family. We refrain from providing the details of this modification.

EXAMPLE 5.3. Consider again the data in Example 4.7. On this data, Algorithm 1 converges in one step and the maximum likelihood distribution is given by a rational function of the data. The corresponding MLEs are

$$\hat{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 \\ 0.5 & 1 & 0.5 & 0.25 \\ 0.25 & 0.5 & 1 & 0.5 \\ 0.125 & 0.25 & 0.5 & 1 \end{bmatrix}, \qquad \hat{J} = \frac{\log(3)}{2} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

This is a very special example, where the following three MLEs all coincide:

1. MLE computed under MTP_2 for general binary distributions.
2. MLE computed for the MTP_2 Ising model over the complete graph.
3. MLE computed for the Ising model over the chain graph in Figure 1.

The equivalence of (2) and (3) follows from Corollary 3.10 whereas (1) and (2) are usually not equivalent.

5.3. *Totally positive Ising models with no external field.* A special example of a symmetric binary distribution is the Ising model with no external field, that is, a family of binary distributions over $\mathcal{X} = \{-1, 1\}^d$ of the form

$$p(x) = \frac{1}{c(J)} \exp\bigl(x^T J x/2\bigr). \tag{5.8}$$

This was termed *the palindromic Ising model* in [28]. The space of canonical parameters is the set $\mathbb{S}_0^d$ of all symmetric $d \times d$ matrices with 0 in the diagonal. The mean parameter is $\Sigma = \Xi = \mathbb{E}XX^T$, which is the correlation matrix because $\Sigma_{ii} = \mathbb{E}X_i^2 = 1$ and $\mathbb{E}X_i = 0$. By Proposition 3.6, the quadratic exponential family is MTP_2 if and only if $J_{ij} \geq 0$ for all $i \neq j$. In [28], these models have been studied as a close proxy to the Gaussian distribution since (5.8) becomes almost identical to the Gaussian density by letting $J = -K$ in this expression.

As a consequence of Proposition 4.15, we note that Algorithm 1 also converges for palindromic Ising models by working with the symmetrized sample $U + U^s$. However, the algorithm can be simplified using

$$\begin{aligned} e_{ij}(1,1) = e_{ij}(-1,-1) &= (1 + M_{ij})/4, \\ e_{ij}(1,-1) = e_{ij}(-1,1) &= (1 - M_{ij})/4. \end{aligned}$$

In addition,

$$\tilde{\Delta}_{ij}(\lambda) = \frac{1}{2} \log \frac{p_{ij}(-1,1)(1 + M_{ij} + \lambda)}{p_{ij}(1,1)(1 - M_{ij} - \lambda)} \tag{5.9}$$

can be used to determine λ to ensure the MTP_2 property is preserved under the update. We refrain from giving the full details of the simplified steps in this algorithm.

6. Application to psychological disorders. In this section, we illustrate the developed methods via a real data case study. We analyze data obtained from the National Comorbidity Survey Replication study [2, 23] (NCS-R data), which was also analyzed in [12]. The data consists of 9282 observations of 18 binary variables, namely `depr` (Depressed mood), `inte` (Loss of interest), `weig` (Weight problems), `mSle` (Sleep problems), `moto` (Psychomotor disturbances), `mFat` (Fatigue), `repr` (Self reproach), `mCon` (Concentration problems), `suic` (Suicidal ideation), `anxi` (Chronic anxiety/worry), `even` (Anxiety $>$ 1 event), `ctrl` (No control over anxiety), `edge` (Feeling on edge), `gFat` (Fatigue), `irri` (Irritable), `gCon` (Concentration problems), `musc` (Muscle tension), `gSle` (Sleep problems). These variables are symptoms related to two disorders, namely major depression (MD) and generalized anxiety disorder (GAD). The symptoms that are known to appear in both disorders are sleep problems, fatigue and concentration problems. These so-called bridge variables appear in pairs `mSle`, `gSle`, `mFat`, `gFat` and `mCon`, `gCon`.

The contingency table resulting from this dataset is very sparse with only 872 out of 65,536 elementary events observed; 5667 out of 9282 respondents recorded none of the listed symptoms. All variables are positively correlated in the sample. Although the sample distribution is not MTP_2, assuming total positivity is justified in this application, since the symptoms are likely to appear jointly. The sample does not satisfy the conditions of Theorem 4.5, because the variables `anxi` and `even` are perfectly correlated with each other and with seven

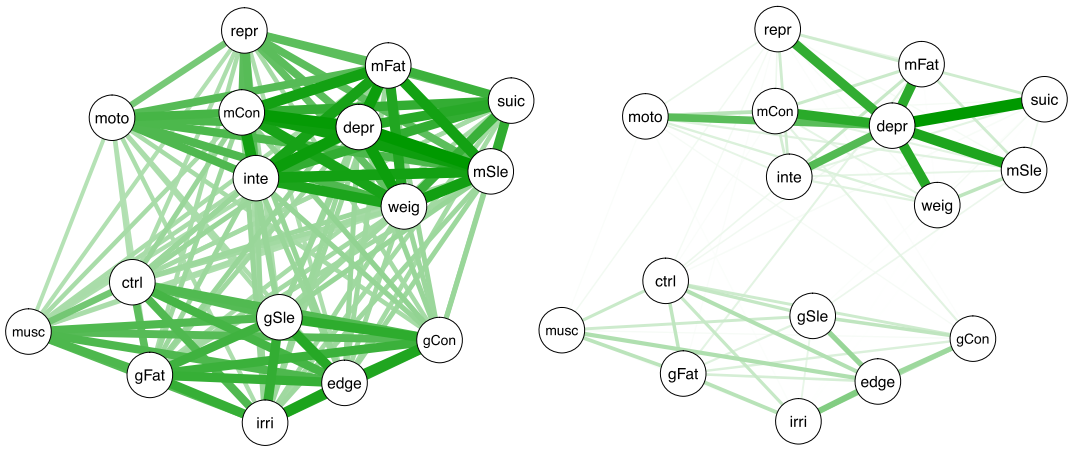


FIG. 2. (*left*) *Sample correlation network for the NCS-R data*, (*right*) *the corresponding network of* $\hat{J}$.

other variables in the sample distribution. For the analysis, we therefore removed these two problematic variables and ran Algorithm 1 on the remaining contingency table of size 2^{16}. We used a convergence criterion of $\epsilon = 10^{-4}$. The algorithm converged after 28 iterations through all 120 variable pairs which in our current rough implementation took 37 minutes on a laptop. We note that using the algorithm and software developed in [9] fitting the unconstrained MTP_2 model failed due to space limitations.

Figure 2 shows the network corresponding to the sample correlation matrix (left) and the MLE $\hat{J}$ (right). The magnitude of an entry ij in the matrix is represented by the thickness of the corresponding edge. The sample correlation network including the two nodes `anxi` and `even` is also shown in Figure 2b of [12]. The sparsity of the MLE $\hat{J}$ as compared to the sample correlation matrix is striking; it contains 72 edges as compared to 120 edges in the complete graph on 16 vertices. In addition, the graphical model given by $\hat{J}$ cleanly separates into two blocks with the upper block prominently containing a star graph with center `depr`. This resembles Figure 4 in [12], where this subgraph is called a causal skeleton of the covariance graph and was obtained based on rankings by 12 clinicians. Moreover, we note that the bottom block, while less prominently, also contains a star graph centered at the variable `edge`. Finally, note that the three most significant edges across the two blocks are between pairs of bridge variables. This analysis shows that Algorithm 1 resulted in an interpretable sparse graphical model with a network that seems relevant for the application.

The graphical model learned by Algorithm 1 fits reasonably well: The value of the log-likelihood function at the MLE is -28,767.3, while the value of the log-likelihood function of the unrestricted Ising model (fitted using the `loglin` function in R) is -28,682.45. This results in a likelihood ratio statistic of 169.7 which appears high compared to a χ^2 distribution with $120 - 72 = 48$ degrees of freedom. However, the exact and asymptotic distributions of this statistic are unknown; the asymptotic distribution is a mixture of χ^2-distributions with different degrees of freedom, but with unknown weights.

We also calculated the split-likelihood ratio test statistic as described in [35] and this resulted in a test statistic of $U_n = 1.8 \times 10^{-58}$ which does not reject the MTP_2 hypothesis for any level α as it should be compared to $1/\alpha$. Hence it appears that the MTP_2 analysis of this dataset is appropriate.

Acknowledgments. We would like to thank Antonio Forcina for making his MATLAB code from [9] available to us. We have also benefited from discussions with Béatrice de Tilière.

Funding. This research was supported through the program "Research in Pairs" by the Mathematisches Forschungsinstitut Oberwolfach in 2018. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM and a Simons Investigator Award.

Piotr Zwiernik was supported by the Spanish Ministry of Economy and Competitiveness (MTM2015-67304-P), Beatriu de Pinós Fellowship (2016 BP 00002) and the program Ayudas Fundación BBVA (2017).

REFERENCES

[1] AGOSTINI, D. and AMÉNDOLA, C. (2019). Discrete Gaussian distributions via theta functions. *SIAM J. Appl. Algebra Geom.* **3** 1–30. MR3904412 https://doi.org/10.1137/18M1164937

[2] ALEGRIA, M., JACKSON, J. S. J. S., KESSLER, R. C. and TAKEUCHI, D. (2016). Collaborative Psychiatric Epidemiology Surveys (CPES), 2001–2003 [United States].

[3] ALLMAN, E. S., BAÑOS, H., EVANS, R., HOŞTEN, S., KUBJAS, K., LEMKE, D., RHODES, J. A. and ZWIERNIK, P. (2019). Maximum likelihood estimation of the latent class model through model boundary decomposition. *J. Algebr. Stat.* **10** 51–84. MR3947124 https://doi.org/10.18409/jas.v10i1.75

[4] ALLMAN, E. S., RHODES, J. A., STURMFELS, B. and ZWIERNIK, P. (2015). Tensors of nonnegative rank two. *Linear Algebra Appl.* **473** 37–53. MR3338324 https://doi.org/10.1016/j.laa.2013.10.046

[5] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *J. Mach. Learn. Res.* **13** 2293–2337. MR2973603

[6] BACH, F. (2013). Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning* **6** 145–373.

[7] BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, Chichester. MR0489333

[8] BARTOLUCCI, F. and BESAG, J. (2002). A recursive algorithm for Markov random fields. *Biometrika* **89** 724–730. MR1929176 https://doi.org/10.1093/biomet/89.3.724

[9] BARTOLUCCI, F. and FORCINA, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *Ann. Statist.* **28** 1206–1218. MR1811325 https://doi.org/10.1214/aos/1015956713

[10] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* 192–236.

[11] BHATTACHARYA, B. (2012). Covariance selection and multivariate dependence. *J. Multivariate Anal.* **106** 212–228. MR2887689 https://doi.org/10.1016/j.jmva.2011.11.002

[12] BORSBOOM, D. and CRAMER, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* **9** 91–121.

[13] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[14] COLANGELO, A., SCARSINI, M. and SHAKED, M. (2005). Some notions of multivariate positive dependence. *Insurance Math. Econom.* **37** 13–26. MR2156593 https://doi.org/10.1016/j.insmatheco.2004.09.004

[15] COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. *Statistics for Engineering and Information Science*. Springer, New York. MR1697175

[16] DJOLONGA, J. and KRAUSE, A. (2015). Scalable variational inference in log-supermodular models. arXiv:1502.06531.

[17] EGILMEZ, H. E., PAVEZ, E. and ORTEGA, A. (2019). Graph learning from filtered signals: Graph system and diffusion kernel identification. *IEEE Trans. Signal Inf. Process. Netw.* **5** 360–374. MR3949864 https://doi.org/10.1109/TSIPN.2018.2872157

[18] FALLAT, S., LAURITZEN, S., SADEGHI, K., UHLER, C., WERMUTH, N. and ZWIERNIK, P. (2017). Total positivity in Markov structures. *Ann. Statist.* **45** 1152–1184. MR3662451 https://doi.org/10.1214/16-AOS1478

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

[20] GEIGER, D., MEEK, C. and STURMFELS, B. (2006). On the toric algebra of graphical models. *Ann. Statist.* **34** 1463–1492. MR2278364 https://doi.org/10.1214/009053606000000263

[21] JIROUŠEK, R. and PŘEUČIL, R. (1995). On the effective implementation of the iterative proportional fitting procedure. *Comput. Statist. Data Anal.* **19** 177–189.

[22] KARLIN, S. and RINOTT, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10** 467–498. MR0599685 https://doi.org/10.1016/0047-259X(80)90065-2

[23] KESSLER, R. C., BERGLUND, P., CHIU, W. T., DEMLER, O., HEERINGA, S., HIRIPI, E., JIN, R., PENNELL, B.-E., WALTERS, E. E. et al. (2004). The US national comorbidity survey replication (NCS-R): Design and field procedures. *Int. J. Methods Psychiatr. Res.* **13** 69–92.

[24] LAURITZEN, S. (2002). Lectures on Contingency Tables, Electronic edition. Earlier editions: 1979, 1982, 1989.

[25] LAURITZEN, S., UHLER, C. and ZWIERNIK, P. (2019). Maximum likelihood estimation in Gaussian models under total positivity. *Ann. Statist.* **47** 1835–1863. MR3953437 https://doi.org/10.1214/17-AOS1668

[26] LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York. MR1419991

[27] MALIOUTOV, D. V., JOHNSON, J. K. and WILLSKY, A. S. (2006). Walk-sums and belief propagation in Gaussian graphical models. *J. Mach. Learn. Res.* **7** 2031–2064. MR2274432

[28] MARCHETTI, G. M. and WERMUTH, N. (2016). Palindromic Bernoulli distributions. *Electron. J. Stat.* **10** 2435–2460. MR3545465 https://doi.org/10.1214/16-EJS1175

[29] MOUSSOURIS, J. (1974). Gibbs and Markov random systems with constraints. *J. Stat. Phys.* **10** 11–33. MR0432132 https://doi.org/10.1007/BF01011714

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

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

[32] ROBEVA, E., STURMFELS, B. and UHLER, C. (2019). Geometry of log-concave density estimation. *Discrete Comput. Geom.* **61** 136–160. MR3925548 https://doi.org/10.1007/s00454-018-0024-y

[33] SLAWSKI, M. and HEIN, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra Appl.* **473** 145–179. MR3338330 https://doi.org/10.1016/j.laa.2014.04.020

[34] STUDENÝ, M. (2005). *Probabilistic Conditional Independence Structures. Information Science and Statistics*. Springer, London. MR3183760

[35] WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference using the split likelihood ratio test. *Proc. Natl. Acad. Sci. USA* **117** 16880–16890.