Maximum likelihood estimation in Gaussian models under total positivity

Lauritzen, Steffen L.; Uhler, Caroline; Zwiernik, Piotr

Published in:
Annals of Statistics

DOI:
http://dx.doi.org/10.1214/17-AOS1668

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

The Annals of Statistics
2019, Vol. 47, No. 4, 1835–1863
https://doi.org/10.1214/17-AOS1668

MAXIMUM LIKELIHOOD ESTIMATION IN GAUSSIAN MODELS UNDER TOTAL POSITIVITY

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

Massachusetts Institute of Technology, University of Copenhagen and Universitat Pompeu Fabra

We analyze the problem of maximum likelihood estimation for Gaussian distributions that are multivariate totally positive of order two (MTP_2). By exploiting connections to phylogenetics and single-linkage clustering, we give a simple proof that the maximum likelihood estimator (MLE) for such distributions exists based on $n \geq 2$ observations, irrespective of the underlying dimension. Slawski and Hein [*Linear Algebra Appl.* **473** (2015) 145–179], who first proved this result, also provided empirical evidence showing that the MTP_2 constraint serves as an implicit regularizer and leads to sparsity in the estimated inverse covariance matrix, determining what we name the ML graph. We show that we can find an upper bound for the ML graph by adding edges corresponding to correlations in excess of those explained by the maximum weight spanning forest of the correlation matrix. Moreover, we provide globally convergent coordinate descent algorithms for calculating the MLE under the MTP_2 constraint which are structurally similar to iterative proportional scaling. We conclude the paper with a discussion of signed MTP_2 distributions.

1. Introduction. Total positivity is a special form of positive dependence between random variables that became an important concept in modern statistics; see, for example, [3, 8, 23]. This property (also called the MTP_2 property) appeared in the study of stochastic orderings, asymptotic statistics and in statistical physics [15, 31]. Families of distributions with this property lead to many computational advantages [2, 11, 33]. In a recent paper [13], the MTP_2 property was studied in the context of graphical models and conditional independence in general. It was shown that MTP_2 distributions have desirable Markov properties. Our paper can be seen as a continuation of this work with a focus on Gaussian distributions.

A p-variate real-valued distribution with density f w.r.t. a product measure μ is *multivariate totally positive of order 2* (MTP_2) if the density satisfies

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y).$$

Received February 2017; revised November 2017.

[1]Supported in part by DARPA Grant W911NF-16-1-0551, NSF Grant DMS-1651995, ONR Grant N00014-17-1-2147 and a Sloan Fellowship.

[2]Supported in part by MINECO (MTM2015-67304-P) and Beatriu de Pinós Fellowship.

MSC2010 subject classifications. Primary 60E15, 62H99; secondary 15B48.

Key words and phrases. MTP_2 distribution, attractive Gaussian Markov random field (GMRF), nonfrustrated GRMF, Gaussian graphical model, inverse M-matrix, ultrametric.

A multivariate Gaussian distribution with mean μ and a positive definite covariance matrix Σ is MTP_2 if and only if the concentration matrix $K := \Sigma^{-1}$ is a symmetric *M-matrix*, that is, $K_{ij} \leq 0$ for all $i \neq j$ or, equivalently, if all partial correlations are nonnegative. Such distributions were considered by Bølviken [5] and Karlin and Rinott [25]. Moreover, Gaussian graphical models, or Gaussian Markov random fields, were studied in the context of totally positive distributions in [29]. MTP_2 Gaussian graphical models were shown to form a sub-class of *non-frustrated* Gaussian graphical models, which themselves are a sub-class of *walk-summable* Gaussian graphical models. Efficient structure estimation algorithms for MTP_2 Gaussian graphical models were given in [1] based on thresholding covariances after conditioning on subsets of variables of limited size. Efficient learning procedures based on convex optimization were suggested by Slawski and Hein [37] and this paper is closely related to their approach; see also [4] and [12].

Throughout this paper, we assume that we are given n i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$, where Σ is an unknown positive definite matrix of size $p \times p$. Without loss of generality, we assume that $\mu = 0$ and we focus on the estimation of Σ. We denote the sample covariance matrix based on n samples by S. Then the log-likelihood function is, up to additive and multiplicative constants, given by

$$\ell(K; S) = \log\det K - \operatorname{tr}(SK). \tag{1}$$

We denote the cone of real symmetric matrices of size $p \times p$ by $\mathbb{S}^p$, its positive definite elements by $\mathbb{S}^p_{\succ 0}$ and its positive semidefinite elements by $\mathbb{S}^p_{\succeq 0}$. Note that $\ell(K; S)$ is a strictly concave function of $K \in \mathbb{S}^p_{\succeq 0}$. Since M-matrices form a convex subset of $\mathbb{S}^p_{\succ 0}$, the optimization problem for computing the *maximum likelihood estimator* (MLE) for MTP_2 Gaussian models is a convex optimization problem. Slawski and Hein [37] showed that the MLE exists with probability one when $n \geq 2$; that is, the global maximum of this optimization problem is attained. This yields a drastic reduction from $n \geq p$ without the MTP_2 constraint. In addition, they provided empirical evidence showing that the MTP_2 constraint serves as an implicit regularizer and leads to sparsity in the concentration matrix K.

In this paper, we analyze the sparsity pattern of the MLE $\hat{K}$ under the MTP_2 constraint. For a $p \times p$ matrix K, we let $G(K)$ denote the undirected graph on p nodes with an edge ij if and only if $K_{ij} \neq 0$. In Proposition 4.3 we obtain a simple upper bound for the ML graph $G(\hat{K})$ by adding edges to the smallest *maximum weight spanning forest* (MWSF) corresponding to empirical correlations in excess of those provided by the MWSF. We illustrate the issues in the following example.

EXAMPLE 1.1. We consider the `carcass` data that are discussed in [19] and can be found in the `R`-library `gRbase`. This data set contains measurements of the thickness of meat and fat layers at different locations on the back of a slaughter pig together with the lean meat percentage on each of 344 carcasses. For our analysis, we ignore the lean meat percentage, since by definition, this variable should

be negatively correlated with fat and positively correlated with meat so the joint distribution is unlikely to be MTP_2. The sample correlation matrix R for these data is

$$R = \begin{array}{c} \begin{matrix} \text{Fat11} & \text{Meat11} & \text{Fat12} & \text{Meat12} & \text{Fat13} & \text{Meat13} \end{matrix} \\ \begin{pmatrix} 1.00 & 0.04 & 0.84 & 0.08 & 0.82 & -0.03 \\ 0.04 & 1.00 & 0.04 & 0.87 & 0.13 & 0.86 \\ 0.84 & 0.04 & 1.00 & 0.01 & 0.83 & -0.03 \\ 0.08 & 0.87 & 0.01 & 1.00 & 0.11 & 0.90 \\ 0.82 & 0.13 & 0.83 & 0.11 & 1.00 & 0.02 \\ -0.03 & 0.86 & -0.03 & 0.90 & 0.02 & 1.00 \end{pmatrix} \end{array} \begin{matrix} \text{Fat11} \\ \text{Meat11} \\ \text{Fat12} \\ \text{Meat12} \\ \text{Fat13} \\ \text{Meat13} \end{matrix}$$

and its inverse, scaled to have diagonal elements equal to one, $\tilde{K}$, is

$$\tilde{K} = \begin{array}{c} \begin{matrix} \text{Fat11} & \text{Meat11} & \text{Fat12} & \text{Meat12} & \text{Fat13} & \text{Meat13} \end{matrix} \\ \begin{pmatrix} 1.00 & 0.16 & -0.52 & -0.31 & -0.40 & 0.19 \\ 0.16 & 1.00 & -0.05 & -0.42 & -0.17 & -0.37 \\ -0.52 & -0.05 & 1.00 & 0.25 & -0.45 & -0.17 \\ -0.31 & -0.42 & 0.25 & 1.00 & -0.02 & -0.61 \\ -0.40 & -0.17 & -0.45 & -0.02 & 1.00 & 0.10 \\ 0.19 & -0.37 & -0.17 & -0.61 & 0.10 & 1.00 \end{pmatrix} \end{array} \begin{matrix} \text{Fat11} \\ \text{Meat11} \\ \text{Fat12} \\ \text{Meat12} \\ \text{Fat13} \\ \text{Meat13} \end{matrix}$$

Note that the off-diagonal entries of $\tilde{K}$ are the negative empirical partial correlations. This sample distribution is not MTP_2; the positive entries in $\tilde{K}$ are highlighted in red. The MLE under MTP_2 can be computed, for example, using `cvx` [17] in `matlab` or using one of the simple coordinate descent algorithms discussed in Section 2. In this particular example, the MLE can also be obtained through the explicit formula (14) in Section 4. The MLE of the correlation matrix, rounded to 2 decimals, is

$$\hat{R} = \begin{array}{c} \begin{matrix} \text{Fat11} & \text{Meat11} & \text{Fat12} & \text{Meat12} & \text{Fat13} & \text{Meat13} \end{matrix} \\ \begin{pmatrix} 1.00 & 0.10 & 0.84 & 0.09 & 0.82 & 0.09 \\ 0.10 & 1.00 & 0.11 & 0.87 & 0.13 & 0.86 \\ 0.84 & 0.11 & 1.00 & 0.09 & 0.83 & 0.09 \\ 0.09 & 0.87 & 0.09 & 1.00 & 0.11 & 0.90 \\ 0.82 & 0.13 & 0.83 & 0.11 & 1.00 & 0.11 \\ 0.09 & 0.86 & 0.09 & 0.90 & 0.11 & 1.00 \end{pmatrix} \end{array} \begin{matrix} \text{Fat11} \\ \text{Meat11} \\ \text{Fat12} \\ \text{Meat12} \\ \text{Fat13} \\ \text{Meat13} \end{matrix}$$

The entries of $\hat{R}$ that changed compared to the sample correlation matrix R are highlighted in blue.[3] The sparsity pattern of $\hat{K} = \hat{\Sigma}^{-1}$ is captured by the ML graph $G(\hat{K})$ shown in Figure 1.

[3]We note that $\hat{\Sigma}_{45} > S_{45}$; the entries appear equal only because of the 2-digit rounding.

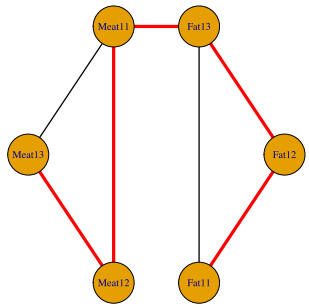


FIG. 1. *Undirected Gaussian graphical model for the* `carcass` *data obtained by estimating under the* MTP_2 *assumption. The thick red edges correspond to the MWSF of the correlation matrix.*

Note that all edges corresponding to blue entries in $\hat{R}$ are missing in this graph. As we show in Proposition 2.2, this is a consequence of the KKT conditions. Consider now the maximum weight spanning forest of the complete graph with weights given by the entries of R. In this example, the spanning forest is a chain represented by the thick red edges in Figure 1. By Corollary 4.7, these edges form a spanning tree of the ML graph $G(\hat{K})$.

Interestingly, applying various methods for model selection such as stepwise AIC, BIC or graphical lasso all yield similar graphs, possibly indicating that the MTP_2 assumption is quite reasonable.

The remainder of this paper is organized as follows: In Section 2, we review the duality theory that is known more generally for regular exponential families and specialize it to MTP_2 Gaussian distributions. This embeds the results by Slawski and Hein [37] into the framework of exponential families and also leads to two related coordinate descent algorithms for computing the MLE, one that acts on the entries of K and one that acts on the entries of Σ. In Section 3, we show how the problem of ML estimation for MTP_2 Gaussian distributions is connected to single-linkage clustering and ultrametrics as studied in phylogenetics. These observations result in a simple proof of the existence of the MLE for $n \geq 2$, a result that was first proven in [37]. Our proof is by constructing a primal and dual feasible point of the convex ML estimation problem for MTP_2 Gaussian models. In Section 4, we investigate the structure of the ML graph $G(\hat{K})$ and give a simple upper bound for it. Finally, in Section 5 we discuss how our results can be generalized to so-called *signed* MTP_2 Gaussian distributions, where the distribution is MTP_2 up to sign changes or, equivalently, $|X|$ is MTP_2. Such distributions were introduced by Karlin and Rinott in [24]. We conclude the paper with a brief discussion of various open problems.

2. Duality theory for ML estimation under MTP_2. We start this section by formally introducing absolutely continuous MTP_2 distributions and then discuss

the duality theory for Gaussian MTP_2 distributions. Let $V := \{1, 2, \ldots, p\}$ be a finite set and let $X = (X_i, i \in V)$ be a random vector with density f w.r.t. Lebesgue measure on the product space $\mathcal{X} = \prod_{i \in V} \mathcal{X}_i$, where $\mathcal{X}_i \subseteq \mathbb{R}$ is the state space of X_i. We define the coordinate-wise minimum and maximum as

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

Then we say that X or the distribution of X is *multivariate totally positive of order two* (MTP_2) if its density function f on $\mathcal{X}$ satisfies

$$f(x) f(y) \leq f(x \wedge y) f(x \vee y) \qquad \text{for all } x, y \in \mathcal{X}. \tag{2}$$

In this paper, we concentrate on the Gaussian setting. It is easy to show that a Gaussian distribution with mean μ and covariance matrix Σ is MTP_2 if and only if $K = \Sigma^{-1}$ is a symmetric *M-matrix*, that is, K is positive definite and:

(i) $K_{ii} > 0$ for all $i \in V$,
(ii) $K_{ij} \leq 0$ for all $i, j \in V$ with $i \neq j$.

Properties of M-matrices were studied by Ostrowski [32] who chose the name to honor H. Minkowski. The connection to multivariate Gaussian distributions was established by Bølviken [5] and Karlin and Rinott [25].

We denote the set of all symmetric M-matrices of size $p \times p$ by $\mathcal{M}^p$. Note that $\mathcal{M}^p$ is a convex cone. In fact, it is obtained by intersecting the positive definite cone $\mathbb{S}^p_{\succ 0}$ with all the coordinate half-spaces

$$\mathcal{H}^p_{ij} = \{X \in \mathbb{S}^p \,|\, X_{ij} \leq 0\}$$

with $i \neq j$. For a convex cone $\mathcal{C}$, we denote its closure by $\overline{\mathcal{C}}$. Then $\overline{\mathcal{M}^p}$ is given by $\mathbb{S}^p_{\succeq 0} \bigcap_{i<j} \mathcal{H}^p_{ij}$ and the ML estimation problem for Gaussian MTP_2 models can be formulated as the following optimization problem:

$$\begin{aligned} &\underset{K}{\text{maximize}} && \log\det(K) - \operatorname{trace}(KS) \\ &\text{subject to} && K \in \mathcal{M}^p \end{aligned} \tag{3}$$

This is a convex optimization problem, since the objective function is concave on $\mathbb{S}^p_{\succeq 0}$.

Next, we introduce a second convex cone $\mathcal{N}^p$ that plays an important role for ML estimation in Gaussian MTP_2 models. To formally define this cone, we introduce two partial orders on matrices. Let A, B be two $p \times p$ matrices. Then $A \geq B$ means that $A_{ij} \geq B_{ij}$ for all $(i, j) \in V \times V$, and $A \succeq B$ means that $A - B \in \mathbb{S}^p_{\succeq 0}$. Then the cone $\mathcal{N}^p$ is defined as the negative closure of $\mathbb{S}^p_{\succ 0}$, that is,

$$\mathcal{N}^p = \big\{X \in \mathbb{S}^p \,|\, \exists Y \in \mathbb{S}^p_{\succ 0} \text{ with } X \leq Y \text{ and } \operatorname{diag}(X) = \operatorname{diag}(Y)\big\}.$$

To simplify notation, we will suppress the dependence on p and write $\mathbb{S}$, $\mathbb{S}_{\succeq 0}$, $\mathbb{S}_{\succ 0}$, $\mathcal{M}$ and $\mathcal{N}$, when the dimension is clear. In the following result, we show that the cones $\mathcal{N}$ and $\mathcal{M}$ are dual to each other.

LEMMA 2.1. *The closure of $\mathcal{N}$ is the dual to the cone of M-matrices $\mathcal{M}$, that is*,

$$\overline{\mathcal{N}} = \{S \in \mathbb{S} | \langle S, K\rangle \geq 0 \text{ for all } K \in \mathcal{M}\}. \tag{4}$$

PROOF. We denote the dual of a convex cone $\mathcal{C}$ by $\mathcal{C}^\vee$. Let $\mathcal{C}_1$, $\mathcal{C}_2$ be two convex cones. Then it is an easy exercise to verify that

$$(\mathcal{C}_1 \cap \mathcal{C}_2)^\vee = \mathcal{C}_1^\vee + \mathcal{C}_2^\vee; \tag{5}$$

here $+$ denotes the Minkowski sum. Note that

$$\mathbb{S}_{\succ 0}^\vee = \mathbb{S}_{\succeq 0} \quad \text{and} \quad \mathcal{H}_{ij}^\vee = \mathcal{H}_{ij}.$$

This completes the proof, since $\mathcal{M} = \mathbb{S}_{\succ 0} \bigcap_{i<j} \mathcal{H}_{ij}$ and (5) can be applied inductively to any finite collection of convex cones. □

Using the cones $\mathcal{M}$ and $\mathcal{N}$, we now determine conditions for existence of the MLE in Gaussian MTP_2 models and give a characterization of the MLE. We say that the MLE does not exist if the likelihood does not attain the global maximum.

PROPOSITION 2.2. *Consider a Gaussian* MTP_2 *model. Then the MLE $\hat{\Sigma}$ (and $\hat{K}$) exists for a given sample covariance matrix S on V if and only if $S \in \mathcal{N}$. It is then equal to the unique element $\hat{\Sigma} \succ 0$ that satisfies the following system of equations and inequalities*:

$$(\hat{\Sigma}^{-1})_{ij} \leq 0 \qquad \text{for all } i \neq j, \tag{6}$$

$$\hat{\Sigma}_{ii} - S_{ii} = 0 \qquad \text{for all } i \in V, \tag{7}$$

$$(\hat{\Sigma}_{ij} - S_{ij}) \geq 0 \qquad \text{for all } i \neq j, \tag{8}$$

$$(\hat{\Sigma}_{ij} - S_{ij})(\hat{\Sigma}^{-1})_{ij} = 0 \qquad \text{for all } i \neq j. \tag{9}$$

PROOF. It is straightforward to compute the dual optimization problem and the KKT conditions. In particular, in [37] it was shown that the dual optimization problem to (3) is given by

$$\begin{aligned} &\underset{\Sigma \succeq 0}{\text{minimize}} && -\log\det(\Sigma) - p \\ &\text{subject to} && \Sigma_{ii} = S_{ii} \qquad \text{for all } i \in V, \\ &&& \Sigma_{ij} \geq S_{ij} \qquad \text{for all } i \neq j. \end{aligned} \tag{10}$$

Note that the identity matrix is a strictly feasible point for (3). Hence, the MLE does not exist if and only if the likelihood is unbounded. Since by Slater's constraint qualification strong duality holds for the optimization problems (3) and (10), the MLE does not exist if and only if $S \notin \mathcal{N}$. □

We note that the conditions in Proposition 2.2 were also derived in [37], save for the explicit identification of the dual cone $\mathcal{N}$.

REMARK 2.3. Proposition 2.2 can easily be extended to provide properties for the existence of the MLE and a characterization of the MLE for Gaussian graphical models under MTP_2. In this case, let $G = (V, E)$ be an undirected graph. Then the primal problem has additional equality constraints, namely $K_{ij} = 0$ for all $ij \notin E$, and hence the inequality constraints in the dual problem are restricted to the entries in E, that is, $\Sigma_{ij} \geq S_{ij}$ for all $ij \in E$. Note that if the MLE of Σ based on S exists in the Gaussian graphical model over G, it also exists in the Gaussian graphical model over G under MTP_2, since without the MTP_2 constraint the MLE needs to satisfy $\hat{\Sigma}_{ij} = S_{ij}$ for all $ij \in E$.

We define the *maximum likelihood graph* (ML graph) $\hat{G}$ to be the graph determined by $\hat{K}$, that is, $\hat{G} = G(\hat{K})$, where $\hat{K} = \hat{\Sigma}^{-1}$ is the MLE of K under MTP_2. We then have the following important corollary of Proposition 2.2.

COROLLARY 2.4. *Consider the Gaussian graphical model determined by* $K_{ij} = 0$ *for* $ij \notin E(\hat{G})$, *where* $\hat{G}$ *is the ML graph under* MTP_2. *Let* $\bar{\Sigma}$ *be the MLE of* Σ *under that Gaussian graphical model* (*without the* MTP_2 *constraint*). *Then* $\hat{\Sigma} = \bar{\Sigma}$.

PROOF. The MLE of Σ under the Gaussian graphical model with graph $\hat{G}$ is the unique element $\bar{\Sigma} \succ 0$ satisfying the following system of equations:

$$\begin{aligned} \bar{\Sigma}_{ii} - S_{ii} &= 0 \qquad \text{for all } i \in V, \\ \bar{\Sigma}_{ij} - S_{ij} &= 0 \qquad \text{for all } ij \in E(\hat{G}), \\ \bigl(\bar{\Sigma}^{-1}\bigr)_{ij} &= 0 \qquad \text{for all } ij \notin E(\hat{G}). \end{aligned}$$

Proposition 2.2 says that also $\hat{\Sigma}$ satisfies these equations, and hence we must have $\bar{\Sigma} = \hat{\Sigma}$. $\square$

Note that this corollary highlights the role of the complementary slackness condition (9) in inducing sparsity of the MTP_2 solution.

We emphasize that the MLE under MTP_2 is equivariant w.r.t. changes of scale so that without loss of generality we can assume that the sample covariance is normalized, that is, $S_{ii} = 1$ or, equivalently, $S = R$, where R is the correlation matrix. For certain of the subsequent developments, this represents a convenient simplification.

LEMMA 2.5. *Let* S *be the sample covariance matrix*, R *the corresponding sample correlation matrix. Denote by* $\hat{\Sigma}^S$ *and* $\hat{\Sigma}^R$ *the MLE in Proposition* 2.2 *based on* S *and* R, *respectively. Then*

$$\hat{\Sigma}^S_{ij} = \sqrt{S_{ii} S_{jj}}\, \hat{\Sigma}^R_{ij} \qquad \textit{for all } i, j \in V.$$

Algorithm 1 Coordinate descent on K

Input: Sample covariance matrix S, and precision ε.
Output: MLE $\hat{K} \in \mathcal{M}$.

1. Let $K^0 := K^1 := (\operatorname{diag}(S))^{-1}$.
2. Cycle through entries $u \neq v$ and solve the following optimization problem:

$$\begin{aligned} &\underset{K \succeq 0}{\text{maximize}} && \log\det(K) - \operatorname{trace}(KS) \\ &\text{subject to} && K_{uv} \leq 0, \\ &&& K_{ij} = K^1_{ij} \qquad \text{for all } ij \in (V \times V) \setminus \{uu, vv, uv\}, \end{aligned}$$

   and update $K^1 = K$.
3. If $\|K^0 - K^1\|_1 < \varepsilon$, set $\hat{K} = K^1$. Otherwise, set $K^0 = K^1$ and return to 2.

PROOF. Denote by D a diagonal matrix such that $D_{ii} = \sqrt{S_{ii}}$ and $S = DRD$. The likelihood function based on S is

$$\log\det K - \operatorname{tr}(SK) = \log\det K - \operatorname{tr}(RDKD).$$

If $K' = DKD$, this can be rewritten as $\log\det K' - \operatorname{tr}(RK') - \sum_i \log S_{ii}$. Therefore, if $\hat{K}^R$ is the maximizer of $\log\det K - \operatorname{tr}(RK)$ under the MTP_2 constraints, then $D^{-1}\hat{K}^R D^{-1}$ is also an M-matrix and the maximizer of $\log\det K - \operatorname{tr}(SK)$. □

We end this section by providing simple coordinate descent algorithms for ML estimation under MTP_2. Although interior point methods run in polynomial time, for very large Gaussian graphical models it is usually more practical to apply coordinate descent algorithms. In Algorithms 1 and 2, we describe two methods for computing the MLE that only use optimization problems of size 2×2 which have a simple and explicit solution, and iteratively update the entries of K, respectively of Σ. Algorithms 1 and 2 are inspired by the corresponding algorithms for Gaussian graphical models; see, for example, [10, 39, 41]. Slawski and Hein [37] also provide a coordinate descent algorithm for estimating covariance matrices under MTP_2. However, their method updates one column/row of Σ at a time.

We first analyze Algorithm 1. Let $A = \{u, v\}$ and $B = V \setminus A$. Then note that the objective function can be written in terms of the 2×2 Schur complement $K' = K_{AA} - K_{AB}K_{BB}^{-1}K_{BA}$, since up to an additive constant

$$\log\det K - \operatorname{trace}(KS) = \log\det K' - \operatorname{trace}\big(K'S_{AA}\big).$$

Algorithm 2 Coordinate descent on Σ

Input: Sample covariance matrix $S \succ 0$, and precision ε.
Output: MLE $\hat{\Sigma}$ with $\hat{\Sigma}^{-1} \in \mathcal{M}$.

1. Let $\Sigma^0 := \Sigma^1 := S$
2. Cycle through entries $u \neq v$ and solve the following optimization problem:

$$\begin{aligned} \underset{\Sigma \succeq 0}{\text{maximize}} \quad & \log\det(\Sigma) \\ \text{subject to} \quad & \Sigma_{uv} \geq S_{uv}, \\ & \Sigma_{ij} = \Sigma^1_{ij} \qquad \text{for all } ij \in (V \times V) \setminus \{uv\}. \end{aligned}$$

   and update $\Sigma^1 = \Sigma$.
3. If $\|\Sigma^0 - \Sigma^1\|_1 < \varepsilon$, set $\hat{\Sigma} = \Sigma^1$. Otherwise, set $\Sigma^0 = \Sigma^1$ and return to 2.

Defining $L := K_{AB}K_{BB}^{-1}K_{BA}$, then the optimization problem in step (2) of Algorithm 1 is equivalent to

$$\begin{aligned} \underset{K' \succeq 0}{\text{maximize}} \quad & \log\det(K') - \operatorname{trace}(K'S_{AA}) \\ \text{subject to} \quad & K'_{12} + L_{12} \leq 0. \end{aligned}$$

The unconstrained optimum to this problem is given by $K' = S_{AA}^{-1}$ and is attained if and only if $(S_{AA}^{-1})_{12} + L_{12} \leq 0$, or equivalently, if and only if

$$L_{12} \leq \frac{S_{uv}}{S_{uu}S_{vv} - S_{uv}^2}.$$

Otherwise, the KKT conditions give that $K'_{12} = -L_{12}$.

Maximizing over the remaining two entries of K' leads to a quadratic equation, which has one feasible solution:

$$\begin{aligned} K'_{11} &= \frac{1 + \sqrt{1 + 4S_{uu}S_{vv}L_{12}^2}}{2S_{uu}}, \\ K'_{22} &= \frac{1 + \sqrt{1 + 4S_{uu}S_{vv}L_{12}^2}}{2S_{vv}}, \qquad K'_{12} = -L_{12}. \end{aligned} \tag{11}$$

Then the solution to the optimization problem in step (2) is given by $K_{AA} = K' + L$.

Dual to this algorithm, one can define an algorithm that iteratively updates the off-diagonal entries of Σ by maximizing the log-likelihood in direction Σ_{uv} and keeping all other entries fixed. This procedure is shown in Algorithm 2. If $p > n$,

S is not positive definite; in this case, we use as starting point the single linkage matrix Z that is defined later in (13).

Similarly as for Algorithm 1, the solution to the optimization problem in step (2) can be given in closed-form. Defining $A = \{u, v\}$, $B = V \setminus A$ and $L = \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$, then analogously as in the derivation above, one can show that the solution to the optimization problem in step (2) of Algorithm 2 is given by

$$\Sigma_{uv} = \max\{S_{uv}, L_{12}\}. \tag{12}$$

We end by proving that Algorithms 1 and 2 indeed converge to the MLE. We here assume that $n \geq 2$ to guarantee existence of the MLE. Note that the suggested starting points for both algorithms can be modified.

PROPOSITION 2.6. *Algorithms 1 and 2 converge to the MLE* $\hat{K} = \hat{\Sigma}^{-1} \in \mathcal{M}$.

PROOF. The convergence to the MLE is immediate for Algorithm 2 because it is a coordinate descent method applied to a smooth and strictly concave function; see, for example, [28]. For Algorithm 1, we use the fact that it is an example of iterative partial maximization. To prove convergence to the MLE, we we will show that the assumptions of Proposition A.3 in [26] hold. The log-likelihood function that we are trying to maximize is strictly concave and so the maximum is unique. Clearly, K is the maximum if and only if it is a fixed point of each update. It only remains to show that updates depend continuously on the previous value. For a given S, fix K and consider a sequence of points K_n converging to K. Denote by $\tilde{K}$ and $\tilde{K}_n$ the corresponding one-step updates. We want to show that $\tilde{K}_n$ also converges to $\tilde{K}$. As above, let $A = \{u, v\}$, $B = V \setminus A$, $K' = K_{AA} - K_{AB}K_{BB}^{-1}K_{BA}$ and $L = K_{AB}K_{BB}^{-1}K_{BA}$. Outside of the block $\tilde{K}_{AA}$, this convergence is trivial; so we focus only on the three entries in $\tilde{K}_{AA}$. The function $L_{12} \mapsto (K'_{11}, K'_{22}, K'_{12})$ is continuous if and only if each coordinate is. It is clear that these functions are continuous if $L_{12} \neq \frac{S_{uv}}{S_{uu}S_{vv}-S_{uv}^2}$. It remains to show that if $L_{12} = \frac{S_{uv}}{S_{uu}S_{vv}-S_{uv}^2}$ the update in (11) gives $K' = S_{AA}^{-1}$, which can be easily checked. □

3. Ultrametric matrices and inverse M-matrices. In this section, we exploit the link to ultrametrics in order to construct an explicit primal and dual feasible point of the maximum likelihood estimation problem.

A nonnegative symmetric matrix U is said to be *ultrametric* if:

(i) $U_{ii} \geq U_{ij}$ for all $i, j \in V$,
(ii) $U_{ij} \geq \min\{U_{ik}, U_{jk}\}$ for all $i, j, k \in V$.

We say that a symmetric matrix is an *inverse M-matrix* if its inverse is an M-matrix. The connection between ultrametrics and M-matrices is established by the following result; see [9], Theorem 3.5.

THEOREM 3.1. *Let U be an ultrametric matrix with strictly positive entries on the diagonal. Then U is nonsingular if and only if no two rows are equal. Moreover, if U is nonsingular then U is an inverse M-matrix.*

The main reason why ultrametric matrices are relevant here is the following construction, which is similar to constructions used in in phylogenetics [34], Section 7.2, and single linkage clustering [16].

Let R be a symmetric $p \times p$ positive semidefinite matrix such that $R_{ii} = 1$ for all $i \in V$. Consider the weighted graph $G^+ = G^+(R)$ over V with an edge between i and j whenever R_{ij} is positive and assign to each edge the corresponding positive weight R_{ij}. Note that G^+ in general does not have to be connected. Define a $p \times p$ matrix Z by setting $Z_{ii} = 1$ for all $i \in V$ and

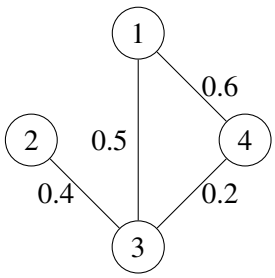
$$Z_{ij} := \max_{P} \min_{uv \in P} R_{uv}, \tag{13}$$

for all $i \neq j$, where the maximum is taken over all paths in G^+ between i and j and is set to zero if no such path exists. We call Z the *single-linkage matrix* based on R.

EXAMPLE 3.2. Suppose that

$$R = \begin{bmatrix} 1 & -0.5 & 0.5 & 0.6 \\ -0.5 & 1 & 0.4 & -0.1 \\ 0.5 & 0.4 & 1 & 0.2 \\ 0.6 & -0.1 & 0.2 & 1 \end{bmatrix}$$

Then G^+ and Z are given by



$$Z = \begin{bmatrix} 1 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1 & 0.4 & 0.4 \\ 0.5 & 0.4 & 1 & 0.5 \\ 0.6 & 0.4 & 0.5 & 1 \end{bmatrix}.$$

For example, to get Z_{12} we consider two paths $1-3-2$ and $1-4-3-2$. The minimum of R_{uv} over the first path is 0.4 and over the second path 0.2. This gives $Z_{12} = 0.4$.

Note that in the above example $Z \geq R$, Z is invertible, and Z^{-1} is an M-matrix. We now show that this is an example of a more general phenomenon.

PROPOSITION 3.3. *Let R be a symmetric $p \times p$ positive semidefinite matrix satisfying $R_{ii} = 1$ for all $i \in V$. Then the single-linkage matrix Z based on R is an ultrametric matrix with $Z_{ij} \geq R_{ij}$ for all $i \neq j$. If, in addition, $R_{ij} < 1$ for all $i \neq j$, then Z is nonsingular and, therefore, an inverse M-matrix.*

PROOF. We first show that Z is an ultrametric matrix. Z is symmetric by definition. Because R is positive semidefinite, $R_{ij} \leq 1$ for all i, j and from (13) it immediately follows that $Z_{ij} \leq 1$ and, therefore, $Z_{ii} \geq Z_{ij}$ for all i, j as needed. Finally, to prove condition (ii) in the definition of ultrametric, let $i, j, k \in V$. Suppose first that i, j, k lie in the same connected component of G^+. Let P_1, P_2 be the paths in G^+ such that $Z_{ik} = \min_{uv \in P_1} R_{uv}$ and $Z_{jk} = \min_{uv \in P_2} R_{uv}$. Let P_{12} be the path between i and j obtained by concatenating P_1 and P_2. Then

$$Z_{ij} = \max_{P} \min_{uv \in P} R_{uv} \geq \min_{uv \in P_{12}} R_{uv} = \min\{Z_{ik}, Z_{jk}\}.$$

Now suppose that i, j, k are not in the same connected component of G^+. In that case, $0 \in \{Z_{ij}, Z_{ik}, Z_{jk}\}$. Because zero is attained at least twice, again $Z_{ij} \geq \min\{Z_{ik}, Z_{jk}\}$. Hence, Z is an ultrametric matrix. The fact that $Z_{ij} \geq R_{ij}$ for all i, j follows directly by noting that the edge ij forms a path between i and j.

Suppose now that $R_{ij} < 1$ for all $i \neq j$. In that case also, $Z_{ij} < 1$ for all $i \neq j$. From this, it immediately follows that no two rows of Z can be equal. Indeed, if the ith row is equal to the jth row for some $i \neq j$, then necessarily $Z_{ij} = Z_{ii} = Z_{jj}$, a contradiction. From Theorem 3.1, it then follows that Z is an inverse M-matrix, which completes the proof. □

As a direct consequence, we obtain the following result.

PROPOSITION 3.4. *Let S be a symmetric positive semidefinite matrix with strictly positive entries on the diagonal and such that $S_{ij} < \sqrt{S_{ii} S_{jj}}$ for all $i \neq j$. Then there exists an inverse M-matrix Z such that $Z \geq S$ and $Z_{ii} = S_{ii}$ for all $i \in V$.*

PROOF. Apply Proposition 3.3 to the normalized version R of S, with entries $R_{ij} := S_{ij}/\sqrt{S_{ii} S_{jj}}$. Because $R_{ij} < 1$ for all $i \neq j$, the corresponding single-linkage matrix Z' is ultrametric with $Z' \geq R$ and Z' is an inverse M-matrix. Define Z by $Z_{ij} = \sqrt{S_{ii} S_{jj}} Z'$. Then $Z \geq S$ and $Z_{ii} = S_{ii}$ for all $i \in V$. Moreover, Z is an inverse M-matrix because Z' is. □

Proposition 3.4 is very important for our considerations. A basic application is an elegant alternative proof of the main result of [37], which says that the MLE under MTP_2 exists with probability one as long as $n \geq 2$. This is in high contrast with the existence of the MLE in Gaussian graphical models without additional constraints; see [40].

THEOREM 3.5 (Slawski and Hein [37]). *Consider a Gaussian* MTP_2 *model and let* S *be the sample covariance matrix. If* $S_{ij} < \sqrt{S_{ii}S_{jj}}$ *for all* $i \neq j$, *then the MLE* $\hat{\Sigma}$ (*and* $\hat{K}$) *exists and it is unique. In particular, if the number* n *of observations satisfies* $n \geq 2$, *then the MLE exists with probability* 1.

PROOF. The sample covariance matrix is a positive semidefinite matrix with strictly positive diagonal entries. We can apply Proposition 3.4 to obtain an inverse M-matrix Z that satisfies $Z \geq S$ and $Z_{ii} = S_{ii}$ for all i. It follows that Z satisfies primal feasibility (6) and dual feasibility (7) and (8). By Proposition 2.2, the MLE exists and it is unique by convexity of the problem. □

REMARK 3.6. Combining this result with Corollary 2.4 we note that the cliques of $\hat{G}$ can at most be of size n. In this way the sparsity of $\hat{G}$ automatically adjusts to the sample size.

The matrix Z can be computed efficiently.[4] To see that, note first that in Example 3.2 we could first consider the chain T of the form $2-3-1-4$, which is the maximal weight spanning forest of G^+ and then construct Z by

$$Z_{ij} = \min_{uv=\overline{ij}} R_{uv},$$

where $\overline{ij}$ denotes the unique path between i and j in T. For example, $Z_{12} = 0.4$, which corresponds to the minimal weight on the path $2-3-1$. This is a general phenomenon.

Suppose again that R is a symmetric $p \times p$ positive semidefinite matrix satisfying $R_{ii} = 1$ for all $i \in V$. Let $\mathrm{MWSF}(R)$ be the set of all minimal maximum weight spanning forests of R. Note that all edge weights of any such forest $F \in \mathrm{MWSF}(R)$ must be positive; hence we must have $F \subseteq G^+$. Also, if R is an empirical correlation matrix, then $\mathrm{MWSF}(R)$ will be a singleton with probability one and in such cases we shall mostly speak of *the* MWSF.

PROPOSITION 3.7. *The single-linkage matrix* Z *as defined in* (13) *is block diagonal with blocks corresponding to the connected components of any* $F \in \mathrm{MWSF}(R)$. *Within each block, all elements are strictly positive and given by*

$$Z_{ij} = \min_{uv\in\overline{ij}} R_{uv},$$

where $\overline{ij}$ *is the unique path between* i *and* j *in a maximal weight spanning tree of* R. *In particular,* $Z_{ij} = R_{ij}$ *for all edges of* $\mathrm{MWSF}(R)$.

[4] In our computations, we used the single-linkage clustering method in R.

PROOF. First suppose that $i, j \in V$ lie in two different components of $F \in \mathrm{MWSF}(R)$. This means that there is no path between i and j in G^+ and so, by definition, $Z_{ij} = 0$. Because $Z_{ij} > 0$ if i, j lie in the same component of F, Z is block diagonal with blocks corresponding to connected components of $\mathrm{MWSF}(R)$.

The rest of the proof is an adaptation of a proof of a related result [34], Proposition 7.2.10. Suppose that $i, j \in V$ lie in the same connected component of F and denote the tree in F corresponding to this component by T. By definition, $Z_{ij} \geq \min_{uv \in \overline{ij}} R_{uv}$. Suppose that $Z_{ij} > \min_{uv \in \overline{ij}} R_{uv}$. We obtain the contradiction by showing that under this assumption T cannot be a maximum weight spanning tree of the corresponding connected component of G^+. Let kl be a minimum weight edge in the unique path between i and j in T. Since $Z_{ij} > R_{kl}$, there exists a path P in G^+ between i and j such that $R_{uv} > R_{kl}$ for every uv in P. Now deleting kl from T partitions the corresponding connected component of G^+ into two sets with i being in one and j being in the other block. Since P connects i and j in G^+, there must be an edge $k'l'$ (distinct from kl) in P whose end vertices lie in different blocks of this partition. Let T' be the spanning tree obtained from T by deleting kl and adding $k'l'$. Since $R_{k'l'} > R_{kl}$, the total weight of T' is greater than T, which is a contradiction. We conclude that $Z_{ij} = \min_{uv \in \overline{ij}} R_{uv}$ for all i, j in the same connected component of G^+. □

To conclude this section, we note that the starting point Σ^0 of Algorithm 2 is arbitrary as long as $\Sigma^0 \succ 0$ and $\Sigma^0 \geq S$. The single-linkage matrix constitutes another generic choice when $S = R$ is used as input. This is a particularly desirable starting point, since it can also be used when $p > n$, in which case $R \notin \mathbb{S}_{\succ 0}$ and hence not feasible.

4. The maximum likelihood graph. Fitting a Gaussian model with MTP_2 constraints tends to induce sparsity in the maximum likelihood estimate $\hat{K}$. In this section, we analyze the sparsity pattern that arises in this way. We assume again without loss of generality that $S = R$ is a sample correlation matrix so that $R_{ii} = 1$ for all i and $R_{ij} < 1$ for all $i \neq j$. Consider again the weighted graph $G^+ = G^+(R)$. We begin this section with a basic lemma that reduces our analysis to the case where the graph G^+ is connected.

LEMMA 4.1. *The MLE* $\hat{\Sigma}$ *under* MTP_2 *is a block diagonal matrix with strictly positive entries in each block. The blocks correspond precisely to trees in* $\mathrm{MWSF}(R)$.

PROOF. First, since $\hat{\Sigma}$ is an inverse M-matrix, it is block diagonal with strictly positive entries in each block; see, for example, Theorem 4.8 in [22]. We will show that each block of $\hat{\Sigma}$ corresponds precisely to a tree in $\mathrm{MWSF}(R)$.

Denote the vertex sets for a forest $F \in \mathrm{MWSF}(R)$ as $T_1, \ldots, T_k$ and the blocks of $\hat{\Sigma}$ as $B_1, \ldots, B_l$. First, for any T_i there must be a j so that $T_i \subseteq B_j$; this is true

since all entries in R along the edges of T_i are positive, and thus $\hat{\Sigma} \geq R > 0$. Thus the block partitioning corresponding to the trees is necessarily finer than that of $\hat{\Sigma}$.

On the other hand, suppose that two different trees T_i and T_j in F are in the same block of $\hat{\Sigma}$ so that $\hat{\Sigma}_{uv} > 0$ for all $u \in T_i$ and $v \in T_j$. Then, as we must have $R_{uv} \leq 0$, also necessarily $\hat{\Sigma}_{uv} - R_{uv} > 0$. Complementary slackness (9) now implies that $\hat{K}_{uv} = 0$ for all $u \in T_i$ and $v \in T_j$, and hence $\hat{K}$ is block-diagonal with blocks corresponding to the trees in F. Since $\hat{\Sigma} = \hat{K}^{-1}$, we also get $\hat{\Sigma}_{uv} = 0$ which contradicts that u and v are in the same block of $\hat{\Sigma}$. $\square$

This result shows that, without loss of generality, we can always assume that G^+ is connected and then $\mathrm{MWSF}(R) = \mathrm{MWST}(R)$ consists of trees only. If there are more than one connected component, we simply compute the MLE for each component separately and combine them together in block diagonal form. Hence, from now on we always assume that all forests in $\mathrm{MWSF}(R)$ are just trees.

4.1. *An upper bound on the ML graph.* In the following, we provide a simple procedure for identifying an upper bound for $\hat{G}$. This procedure relies on the estimation of the standard Gaussian graphical model over the tree $\mathrm{MWSF}(R)$. The MLE under this assumption, denoted by $\tilde{\Sigma}$, can be computed efficiently and it satisfies

$$\tilde{\Sigma}_{ij} = \prod_{uv \in \overline{ij}} R_{uv},$$

where $\overline{ij}$ denotes the unique path between i and j in $\mathrm{MWSF}(R)$; see, for example, [42], Section 8.2. To provide an upper bound on $\hat{G}$, we will make use of a connection to so-called path product matrices: A nonnegative matrix R is a *path product matrix* if for any $i, j \in V$, $k \in \mathbb{N}$, and $1 \leq i_1, \ldots, i_k \leq p$,

$$R_{ij} \geq R_{ii_1} R_{i_1 i_2} \cdots R_{i_k j}.$$

If in addition the inequality is strict for $i = j$, we say that R is a *strict path product matrix*. We note the following.

THEOREM 4.2 (Theorem 3.1, [21]). *Every inverse M-matrix is a strict path product matrix.*

We are now able to provide an upper bound for the ML graph $\hat{G}$.

PROPOSITION 4.3. *The pair ij forms an edge in the ML graph only if*

$$R_{ij} \geq \prod_{uv \in P} R_{uv}$$

for any path P in G^+ between i and j. In particular, $R_{ij} \leq 0$ implies that ij is not an edge of the ML graph.

PROOF. Because $\hat{\Sigma}$ is an inverse M-matrix it is necessarily a path product matrix by Theorem 4.2. In particular, for all i, j and any path P between them,

$$\hat{\Sigma}_{ij} \geq \prod_{uv \in P} \hat{\Sigma}_{uv}.$$

By Proposition 2.2, we also have $\hat{\Sigma}_{uv} \geq R_{uv}$. Thus, if $ij \in \hat{G}$ and P is a path in G^+ we have

$$R_{ij} = \hat{\Sigma}_{ij} \geq \prod_{uv \in P} \hat{\Sigma}_{uv} \geq \prod_{uv \in P} R_{uv}$$

as desired. □

Motivated by this result, we define the *excess correlation graph* (EC graph) $\mathrm{EC}(R)$ of R by the condition

$$i \sim j \quad \Longleftrightarrow \quad R_{ij} \geq \tilde{\Sigma}_{ij} = \prod_{uv \in \overline{ij}} R_{uv}.$$

Thus the EC graph has edges ij whenever the observed correlation between i and j is in excess of or equal to what is explained by the spanning forest; by construction,

$$G(\hat{K}) \subseteq \mathrm{EC}(R).$$

The inclusion is typically strict. For example, if R is an inverse M-matrix, then $\mathrm{EC}(R)$ is the complete graph, whereas $G(\hat{K})$ can be arbitrary; this follows from [13], Proposition 6.3.

4.2. *Some exact results on the ML graph.* Next, we analyze generalization of trees known as block graphs, where edges are replaced by cliques, and give a condition under which the maximum likelihood estimator admits a simple closed-form solution. More formally, G is a *block graph* if it is a chordal graph with only singleton separators. It is natural to study block graphs, since viewing the MLE $\hat{\Sigma}$ as a completion of S, block graphs play the same role for inverse M-matrices as chordal graphs play for Gaussian graphical models; see, for example, [20] and Corollary 7.3 of [13].

We first define a matrix $W = W(R)$ by

$$W_{ij} := \max_{P} \prod_{uv \in P} R_{uv}, \tag{14}$$

where, like in (13), the maximum is taken over all paths in G^+ between i and j and is set to zero if no such path exists. Transforming $D_{ij} = -\log R_{ij}$ gives a distance based interpretation, in which W_{ij} is related to the shortest distance between i and j in G^+ with edge lengths given by D_{uv}. We also have the following simple lemma.

LEMMA 4.4. *The matrix W is a path product matrix. Further, R is a path product matrix if and only if $W(R) = R$.*

PROOF. This is immediate from the definition of W. □

It is easy to show that $Z \geq W \geq R$ and that W is always equal to the MLE $\hat{\Sigma}$ in the case when $p \leq 3$. For general p we do not know conditions on R that assure that W is an inverse M-matrix, or the MLE. Indeed, Example 3.4 in [21] gives a strict path product correlation matrix R, and thus $W = R$, which is not an inverse M-matrix, and thus $W \neq \hat{\Sigma}$. We note that $W = \hat{\Sigma}$ for the `carcass` data discussed in Example 1.1 and, as we shall see in the following, it reflects that in this example, the ML graph is a block graph.

Let $G_R(W)$ be the graph having edges ij exactly when $R_{ij} = W_{ij}$ and no edges otherwise. We then obtain the following result.

PROPOSITION 4.5. *If $G_R(W)$ is a block graph and blocks of W corresponding to cliques are inverse M-matrices, then $\hat{\Sigma} = W$ and $\hat{G} \subseteq G_R(W)$.*

PROOF. Note first that if $\hat{\Sigma} = W$, the KKT conditions (9) imply that $\hat{G} \subseteq G_R(W)$. Let $\tilde{\Sigma}$ denote the maximum likelihood estimate of Σ under the Gaussian graphical model with graph $G_R(W)$. Then, since $G_R(W)$ is a block graph, it follows from [26], equation (5.46) on page 145, that $\tilde{\Sigma}$ is an inverse M-matrix which coincides with W and R on all edges of $G_R(W)$. Thus, $\tilde{\Sigma} = \hat{\Sigma}$ and to show that $\hat{\Sigma} = W$ we just need to argue that $\tilde{\Sigma} = W$.

We proceed by induction on the number $m = |\mathcal{C}|$ of cliques of $G_R(W)$. If there is only one clique in $G_R(W)$, we have $\tilde{\Sigma} = R$ and R is an inverse M-matrix, and hence $\hat{\Sigma} = R = W$. Assume now that the statement holds for $|\mathcal{C}| \leq m$ and assume $G_R(W)$ has $m+1$ cliques. Since $G_R(W)$ is a block graph, there is a decomposition (A, B, S) of $G_R(W)$ into block graphs with at most m cliques and with the separator $S = \{s\}$ being a singleton. But for a decomposition of $G_R(W)$ as above we have from [26], equation (5.31) in Proposition 5.6, and the inductive assumption that

$$\tilde{\Sigma}_{A\cup S} = \hat{\Sigma}_{A\cup S} = W(R_{A\cup S}), \qquad \tilde{\Sigma}_{B\cup S} = \hat{\Sigma}_{B\cup S} = W(R_{B\cup S}).$$

Now let P^* be the path in G^+ such that $W_{ij} = \prod_{uv\in P^*} R_{uv}$ for any two vertices i, j. We claim that all edges in P^* must be edges of $G_R(W)$. Otherwise, suppose P^* contains an edge uv which is not an edge in $G_R(W)$; then $W_{uv} > R_{uv}$ and so if we replace the edge uv with the path realizing W_{uv} the product would be strictly increased, which contradicts the optimality of P^*. Since S is a singleton separator, this also implies that P^* passes through S whenever it involves vertices from both A and B. Suppose that $i, j \in A \cup S$. Then optimality of P^* implies that P^* is contained in $A \cup S$ and so $\tilde{\Sigma}_{A\cup S} = W(R_{A\cup S}) = W_{A\cup S}$ and by the same

argument $\tilde{\Sigma}_{B\cup S} = W_{B\cup S}$. Moreover, if $i \in A$ and $j \in B$ then $W_{ij} = W_{is}W_{sj}$. Now the inductive assumption in combination with the expression [26], p. 140, yields that

$$\tilde{\Sigma}_{ij} = \tilde{\Sigma}_{is}\tilde{\Sigma}_{sj} = W_{is}W_{sj} = W_{ij} \qquad \text{for } i \in A,\ j \in B,$$

and thus $\tilde{\Sigma} = \hat{\Sigma} = W$ as required. □

REMARK 4.6. We note that with probability one, the slackness constraints in (9) are not simultaneously active, and hence in Proposition 4.5 we have almost sure equality between $G_R(W)$ and $\hat{G}$. Thus we can identify $\hat{G}$ without first calculating $\hat{K}$.

We further have the following corollary.

COROLLARY 4.7. *Under the same conditions as in Proposition 4.5 we have* $\mathrm{MWSF}(R) \subseteq \hat{G} \subseteq G_R(W)$.

PROOF. Consider an edge ij between vertices in different cliques of $G_R(W)$ and assume $S_1 = \{s_1\}$ and $S_2 = \{s_2\}$ are (i, j)-separators with $i \sim s_1$ and $j \sim s_2$. Then, since $\hat{G} \subseteq G_R(W)$ we have $i \perp\!\!\!\perp j \,|\, s_1$ and $i \perp\!\!\!\perp j \,|\, s_2$ according to $\hat{\Sigma}$ and therefore

$$\begin{aligned} R_{ij} \le \hat{\Sigma}_{ij} &= \hat{\Sigma}_{is_1}\hat{\Sigma}_{js_1} = \hat{\Sigma}_{is_2}\hat{\Sigma}_{js_2} \\ &= R_{is_1}\hat{\Sigma}_{js_1} = \hat{\Sigma}_{is_2}R_{js_2} < \min\{R_{is_1}, R_{js_2}\}, \end{aligned}$$

so the edge ij can never be part of a MWSF because removing the edge ij would render either s_1 disconnected from i or s_2 disconnected from j and then the weight of the MWSF would increase when replacing ij with is_1 or js_2, respectively. This completes the proof. □

It is not correct in general that $\mathrm{MWSF}(R) \subseteq \hat{G}$ as demonstrated in the following example; although this has been the case in all nonconstructed examples we have considered including the relatively large Example 5.8 below.

EXAMPLE 4.8. The following M-matrix

$$K = \begin{pmatrix} 1 & -0.116 & 0 & 0 & -0.433 \\ -0.116 & 1 & -0.097 & -0.034 & 0 \\ 0 & -0.097 & 1 & -0.149 & -0.413 \\ 0 & -0.034 & -0.149 & 1 & -0.604 \\ -0.433 & 0 & -0.413 & -0.604 & 1 \end{pmatrix}$$

is the inverse of the following correlation matrix

$$R = \begin{pmatrix} 1 & 0.2861 & 0.5745 & 0.6242 & 0.7299 \\ 0.2861 & 1 & 0.2864 & 0.2696 & 0.2872 \\ 0.5745 & 0.2864 & 1 & 0.7149 & 0.7800 \\ 0.6242 & 0.2696 & 0.7149 & 1 & 0.8523 \\ 0.7299 & 0.2872 & 0.7800 & 0.8523 & 1 \end{pmatrix}.$$

Here, $\mathrm{MWSF}(R)$ is the star graph with 5 as its center, but the edge $2 \sim 5$ is not in $G(K)$. Note that all edges in G^+ adjacent to 2 have almost the same weight. We have calculated K^{-1} using rational arithmetic to ensure the phenomenon cannot be explained by rounding error.

5. Gaussian signed MTP_2 distributions. In this section, we discuss how our results can be generalized to so-called signed MTP_2 Gaussian distributions, where the distribution is MTP_2 up to sign swapping. Such distributions were discussed by Karlin and Rinott [24]. More precisely, a random variable X has a *signed* MTP_2 *distribution* if there exists a diagonal matrix D with $D_{ii} = \pm 1$ (called *sign matrix*) such that DX is MTP_2. The following characterization of signed MTP_2 Gaussian distributions is a direct consequence of [24], Theorem 3.1 and Remark 1.3.

PROPOSITION 5.1. *A Gaussian random variable X has a signed* MTP_2 *distribution if and only if $|X|$ is* MTP_2.

Gaussian graphical models with signed MTP_2 distributions are called *nonfrustrated* in the machine learning community. The following result is implicitly stated in [29].

THEOREM 5.2. *A Gaussian random variable X with concentration matrix K has a signed* MTP_2 *distribution if and only if it holds for every cycle $(i_1, \dots, i_k, i_1)$ in the graph $G(K)$ that*

$$(-1)^k K_{i_1 i_2} K_{i_2 i_3} \cdots K_{i_k i_1} > 0. \tag{15}$$

PROOF. The "only if" direction is easy to check. Note that (15) can be rephrased by saying that each cycle in the graph with edge weights given by the off-diagonal entries of $-K$ has an even number of negative edges. The "if" direction can now be recovered from the proof of [29], Corollary 3. □

Signed MTP_2 distributions are relevant, for example, because of their appearance when studying tree models.

PROPOSITION 5.3. *Every Gaussian graphical model over a tree consists of signed* MTP_2 *distributions. The* MTP_2 *distributions among those are precisely those without negative entries in the covariance matrix Σ.*

PROOF. Let T be a tree and $K = \Sigma^{-1}$ be a concentration matrix in the Gaussian graphical model over T. Then $G(K)$ is a subgraph of T and in particular it has no cycles. Hence by Theorem 5.2 it is signed MTP_2. The second part of the statement follows from [13], Corollary 7.3. $\square$

Because signed MTP_2 distributions are closed under taking margins, Proposition 5.3 can be further generalized. The following theorem covers, in particular, Examples 4.1–4.5 in [24].

THEOREM 5.4. *Every distribution on a Gaussian tree model with hidden variables is signed* MTP_2.

Gaussian tree models with hidden variables have many applications, in particular related to modeling evolutionary processes; see, for example, [7, 36]. As an important submodel they contain the Brownian motion tree model [14]. Another example of a Gaussian tree model is the factor analysis model with a single factor; it corresponds to a Gaussian model on a star tree, whose inner node is hidden. The MTP_2 distributions in this model correspond to the distributions in a Spearman model [27, 38], where the hidden factor is interpreted as intelligence.

Let R be a sample correlation matrix. Maximizing the likelihood over all signed MTP_2 Gaussian distributions requires determining the sign matrix D, with $D_{ii} = \pm 1$, that maximizes the likelihood for all 2^p possible matrices DRD. A natural heuristic is to choose D^* such that $D^*_{ii} D^*_{jj} R_{ij} \geq 0$ for all edges ij of MWSF($|R|$), where $|R|$ denotes the matrix whose entries are the absolute values of the entries of R. We provide conditions under which this procedure indeed leads to the MLE under signed MTP_2, and we also provide examples showing that this is not true in general. Quite interestingly, balanced graphs again play an important role in this part of the theory.

First, we describe how to obtain a sign swapping matrix D^* such that $D^*_{ii} D^*_{jj} R_{ij} \geq 0$ for all edges ij of MWSF($|R|$). Root MWSF($|R|$) at node 1, that is, regard MWSF($|R|$) as a directed tree with all edges directed away from 1. Set $D^*_{11} = 1$. Then proceed recursively. For any edge $i \to j$, suppose that D^*_{ii} is known and set $D^*_{jj} := \mathrm{sgn}(D^*_{ii} R_{ij})$. Note that by construction

$$D^*_{ii} := \mathrm{sgn}(R_{1i_1} R_{i_1 i_2} \cdots R_{i_k i}), \tag{16}$$

where $1 \to i_1 \to i_2 \to \cdots \to i_k \to i$ is the unique path from 1 to i in MWSF($|R|$). We set $D^*_{ii} = 0$ if no such path exists. It is easy to check that the resulting D^* satisfies $D^*_{ii} D^*_{jj} R_{ij} \geq 0$ for all edges ij of MWSF($|R|$).

PROPOSITION 5.5. *Suppose that R is a sample correlation matrix whose graph is balanced, that is, such that for every cycle $(i_1, i_2, \ldots, i_k, i_1)$ in the graph $G(R)$:*

$$R_{i_1 i_2} R_{i_2 i_3} \cdots R_{i_k i_1} > 0. \tag{17}$$

Then the MLE based on R over signed MTP_2 *Gaussian distributions is equal to the MLE based on the sample correlation matrix D^*RD^* over* MTP_2 *distributions.*

PROOF. We first show that D^*RD^* has only positive entries. Let i, j be any two nodes and let $1 \to i_1 \to \cdots \to i_k \to i$ and $1 \to j_1 \to \cdots \to j_l \to j$ be the paths in $\mathrm{MWSF}(|R|)$ from 1 to i and j, respectively. By (16), we obtain

$$\operatorname{sgn}\bigl(D^*_{ii}D^*_{jj}R_{ij}\bigr) = \operatorname{sgn}(R_{1i_1}\cdots R_{i_ki}R_{ij}R_{jj_l}\cdots R_{j_11}),$$

which is positive by (17). This shows that without loss of generality we can assume that all entries of R are nonnegative, and hence that D^* is the identity matrix $\mathbb{I}_p$. We now show that the likelihood over MTP_2 distributions given the sample correlation matrix DRD is maximized by $D = \mathbb{I}_p$. This is because $(D_{ii}D_{jj} - 1) \le 0$ and $R_{ij}K_{ij} \le 0$, and hence

$$\ell(K; R) - \ell(K; DRD) = \operatorname{tr}(DRDK) - \operatorname{tr}(RK) = \sum_{i,j}(D_{ii}D_{jj} - 1)R_{ij}K_{ij} \ge 0,$$

which completes the proof. □

Note that any spanning tree T of $G^+(|R|)$ would suffice to identify the sign switches as above.

Proposition 5.5 provides a sufficient condition for D^* to be the optimal sign-switching matrix; that is, it provides a sufficient condition such that for every $K \in \mathbb{S}_{\succ 0}$ and every sign matrix D it holds that

$$\ell\bigl(K; D^*RD^*\bigr) \ge \ell(K; DRD).$$

As a consequence of Proposition 5.5 we obtain the following result for the case when the sample size is 2.

COROLLARY 5.6. *If the sample correlation matrix R is based on $n = 2$ observations, then the MLE over signed* MTP_2 *Gaussian distributions given R is equal to the MLE over* MTP_2 *Gaussian distributions given the modified sample correlation matrix D^*RD^*.*

Note that the case $n = 2$ is special and Proposition 5.5 does not extend to arbitrary sample correlation matrices. In the following, we give a simple counterexample.

EXAMPLE 5.7. Suppose that the sample correlation matrix is

$$R = \begin{bmatrix} 1 & 0.3 & 0.11 & 0.3 \\ 0.3 & 1 & -0.1 & -0.1 \\ 0.11 & -0.1 & 1 & -0.1 \\ 0.3 & -0.1 & -0.1 & 1 \end{bmatrix}.$$

Then MWSF($|R|$) is given by the star graph with edges $1-2$, $1-3$, $1-4$. Since R is positive on these entries, $D^* = \mathbb{I}_p$. But one can check that the corresponding MLE has a lower likelihood than the MLE after changing the sign of the third variable.

The intuition is the following. The log-likelihood based on R is up to an additive constant given by

$$\begin{aligned} \underset{\Sigma}{\text{minimize}} \quad & -\log\det(\Sigma) \\ \text{subject to} \quad & \Sigma_{11} = \Sigma_{22} = \Sigma_{33} = \Sigma_{44} = 1, \\ & \Sigma_{12} \geq R_{12}, \qquad \Sigma_{13} \geq R_{13}, \qquad \Sigma_{14} \geq R_{14}, \\ & \Sigma_{23} \geq 0, \qquad \Sigma_{24} \geq 0, \qquad \Sigma_{34} \geq 0, \\ & \Sigma \succeq 0. \end{aligned}$$

By changing the sign of the third variable, we replace the constraint $1-3$ by two constraints $2-3$ and $3-4$. The resulting optimization problem is

$$\begin{aligned} \underset{\Sigma}{\text{minimize}} \quad & -\log\det(\Sigma) \\ \text{subject to} \quad & \Sigma_{11} = \Sigma_{22} = \Sigma_{33} = \Sigma_{44} = 1, \\ & \Sigma_{12} \geq R_{12}, \qquad \Sigma_{14} \geq R_{14}, \qquad \Sigma_{23} \geq -R_{23}, \qquad \Sigma_{34} \geq -R_{34}, \\ & \Sigma_{13} \geq 0, \qquad \Sigma_{24} \geq 0, \\ & \Sigma \succeq 0. \end{aligned}$$

Note that R_{13} is only slightly larger than $-R_{23}$ and $-R_{24}$. Hence, in essence we are increasing the number of constraints by one, which explains the decrease of the log-likelihood value.

We conclude this paper by illustrating how our results can be applied to factor analysis in psychometrics.

EXAMPLE 5.8. Single factor models are routinely used to study the personalities in psychometrics. Consider the following example from [30]:[5] 240 individuals were asked to rate themselves on the scale 1–9 with respect to 32 different personality traits. The resulting correlation matrix is shown in Figure 2. It appears to have a block structure with predominantly positive entries in each diagonal block and negative entries in the off-diagonal block. Also analyzing the respective variables, they seem to correspond to positive and negative traits. It is therefore natural to

[5]We downloaded the data from http://web.stanford.edu/class/psych253/tutorials/FactorAnalysis.html.

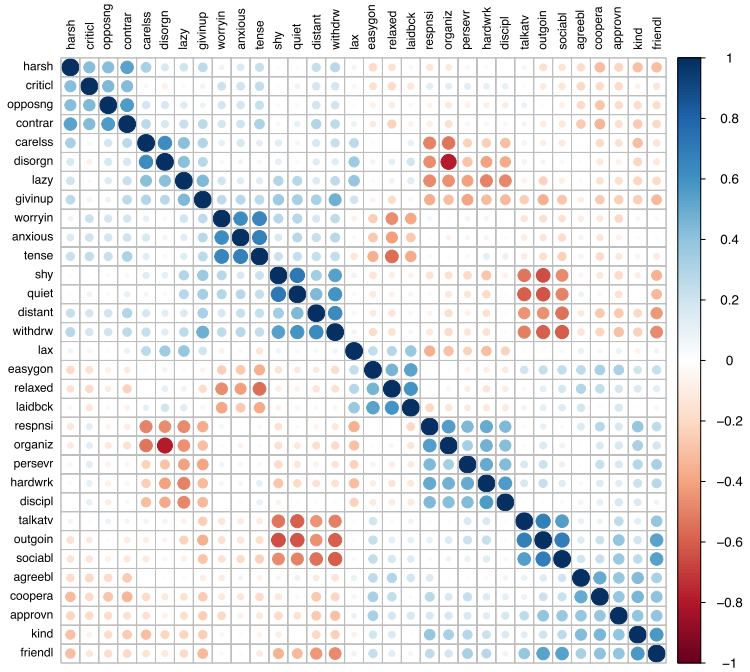


FIG. 2. *Correlation matrix of personality traits from the data set described in* [30].

assume that this data set follows a signed MTP_2 distribution and analyze it under this constraint.

The correlation matrix resulting from the sign switching procedure described in (16) is shown on the left in Figure 3, while the correlation matrix resulting from switching the signs of the 16 (negative) traits that constitute the first block of variables in Figure 2 is shown on the right in Figure 3. These plots suggest that the matrix on the right is closer to being MTP_2. In fact, its log-likelihood [i.e., the value of $\frac{n}{2}(\log\det K - \mathrm{tr}(SK))$] is -2046.146, as compared to the log-likelihood value of -2071.717 resulting from the sign switching procedure described in (16). For comparison, the value of the unconstrained log-likelihood is -1725.075 and the value of the log-likelihood under MTP_2 without sign switching is -2356.639. The unconstrained log-likelihood gives a lower bound of 642.142 on the likelihood ratio statistic to test signed MTP_2 constraints, while the likelihood ratio statistic to test MTP_2 constraints against the saturated model is equal to 1263.128.

The graphical models based on no sign switching and switching the signs of the 16 negative traits are shown in Figure 4. The vertex labels are as shown in Table 1.

The red edges correspond to the maximum weight spanning trees. Red and blue edges together form the edge set of the ML graph so in both of these cases we have $\mathrm{MWSF}(R) \subset \hat{G}$. Finally, the grey edges are the remaining edges in the EC graph. As expected, the graph on the right looks denser. The interpretation of the

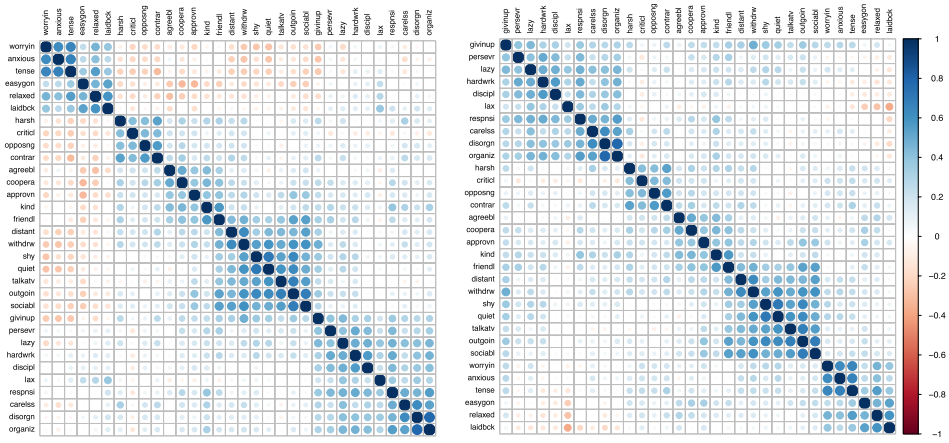


FIG. 3. *The correlation matrix of the data set on personality traits after performing the sign switches as defined in* (16) *is shown on the left. The correlation matrix resulting from switching the signs of the* 16 (*negative*) *traits that constitute the first block of variables in Figure* 2 *is shown on the right.*

spanning tree in both cases is very different. Edges in the first one connect similar personalities such as 6–24 (agreeable and cooperative), 12–22 (outgoing and sociable), 11–23 (disorganized and lazy). On the other hand, the second tree looks similar but it links also some almost perfect opposite personalities such as 12–14 (outgoing and shy), 22–30 (sociable and withdrawn), 11–26 (disorganized and or-

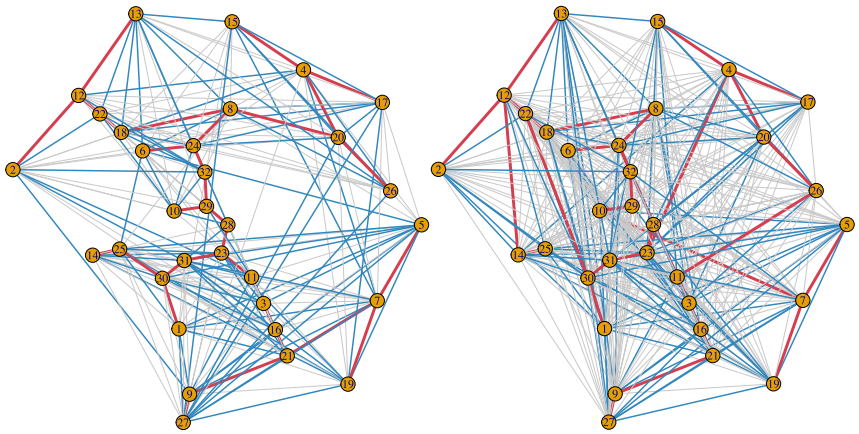


FIG. 4. *On the left, the graphical models resulting from estimation under* MTP_2 *based on the correlation matrix shown in Figure* 2 *and, on the right, the correlation matrix shown in Figure* 3 (*right*). *The thin gray edges correspond to the edges of the EC graph that are not part of the ML graph. The blue edges represent edges of the ML graph that are not part of the minimum weight spanning tree. The latter is represented by thick red edges.*

TABLE 1
Vertex labeling for Figure 4

1	2	3	4	5	6	7	8
distant	talkatv	carelss	hardwrk	anxious	agreebl	tense	kind
9	10	11	12	13	14	15	16
opposng	relaxed	disorgn	outgoin	approvn	shy	discipl	harsh
17	18	19	20	21	22	23	24
persevr	friendl	worryin	respnsi	contrar	sociabl	lazy	coopera
25	26	27	28	29	30	31	32
quiet	organiz	criticl	lax	laidbck	withdrw	givinup	easygon

ganized), 7–10 (tense and relaxed). Note that none of these four edges are part of the ML graph on the left in Figure 4.

6. Discussion. In this article, we have investigated maximum likelihood estimation for Gaussian distributions under the restriction of multivariate total positivity, used a connection to ultrametrics to show that it has a unique solution when the number of observations is at least two, shown that under certain circumstances the MLE can be obtained explicitly, and given convergent algorithms for calculating the MLE. For signed MTP_2 distributions, we have also given conditions under which a heuristic procedure for applying sign changes is correct and can be used to obtain the MLE.

It remains an issue to consider the asymptotic properties of the estimators we have given, and to derive reliable methods for identifying whether a given sample is consistent with the MTP_2 assumption.

On the former issue, standard arguments for convex exponential families ensure that if the true value K_0 is an M-matrix, $\hat{K}$ is a consistent estimator of K_0; and this is true whether or not the MTP_2 assumption is invoked.

Another question is whether the ML graph $\hat{G}$ will be consistent for the true dependence graph. It is clear that without some form of penalty or thresholding, it cannot be the case. For example, if $p = 2$ and the true Σ is a diagonal matrix, the distribution of the empirical correlation R_{12} will be symmetric around 0. Hence, with probability $1/2$ the ML graph contains an edge between 1 and 2 and with probability $1/2$ it does not contain such an edge. This phenomenon persists for any number of observations n. Thus, to achieve consistent estimation of the dependence graph of Σ, some form of penalty for complexity or thresholding must be applied, the latter being suggested by [37], who also suggest a refitting after thresholding to ensure positive definiteness of the thresholded matrix. However, positive definiteness is automatically ensured, as shown below.

PROPOSITION 6.1. *Let K be an M-matrix over V and $G = (V, E)$ an undirected graph. Define K^G by*

$$K^G_{uv} = \begin{cases} K_{uv} & \text{if } u = v \text{ or } uv \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Then K^G is an M-matrix.

PROOF. We may, without loss of generality, assume that K is scaled such that all diagonal elements are equal to 1; also it is clearly sufficient to consider the case when only a single off-diagonal entry K_{uv} is replaced by zero. We have to show that the resulting matrix K^G is positive definite.

Now, let $A = \{u, v\}$ and $B = V \setminus A$ and consider the Schur complements

$$K/K_{BB} = K_{AA} - K_{AB}(K_{BB})^{-1}K_{BA};$$
$$K^G/K_{BB} = K^G_{AA} - K_{AB}(K_{BB})^{-1}K_{BA}.$$

Since $K^G_{BB} = K_{BB}$, K^G is positive definite if and only if K^G/K_{BB} is. Because K is an M-matrix, all entries in $K_{AB}(K_{BB})^{-1}K_{BA}$ are nonnegative. Hence, we can write the Schur complements as

$$K/K_{BB} = \begin{pmatrix} 1-c & -(a+b) \\ -(a+b) & 1-d \end{pmatrix}; \qquad K^G/K_{BB} = \begin{pmatrix} 1-c & -b \\ -b & 1-d \end{pmatrix},$$

where $c, d \in (0, 1)$ and $a, b \geq 0$. Since K is positive definite, we have

$$(a+b)^2 < (1-c)(1-d)$$

and hence

$$b^2 < (1-c)(1-d) - a^2 - 2ab \leq (1-c)(1-d)$$

implying that K^G/K_{BB} is positive definite. This completes the proof. □

The consistency of the estimator $\hat{K}$ ensures that the ML graph will eventually contain the true dependence graph when n becomes large and with an appropriate thresholding or penalization; this ensures that the true graph can be recovered, as also argued in [37].

The issue of the asymptotic distribution of the likelihood ratio test for MTP_2 is an instance of testing a convex hypothesis within an exponential family of distributions. In our particular case, the convex hypothesis is a polyhedral cone with facets determined by the dependence graph $G(K)$. In such cases, the likelihood ratio test for the convex hypothesis typically has an asymptotic distribution which is a mixture of χ^2-distributions with degrees of freedom determined by the co-dimension of these facets; see, for example, the analysis of the case of multivariate positivity in models for binary data by [3], using results of [35].

While these issues are both interesting and important, we consider them to be outside the scope of the present paper as they may be most efficiently dealt with in the more general context of exponential families, containing both the Gaussian and binary cases as special instances. We plan to return to these and other problems in the future.

Acknowledgments. We thank two anonymous referees for their helpful comments.

REFERENCES

[1] 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

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

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

[4] BHATTACHARYA, B. (2012). Covariance selection and multivariate dependence. *J. Multivariate Anal.* **106** 212–228.

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

[6] BUHL, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Stat.* **20** 263–270.

[7] CHOI, M. J., TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning latent tree graphical models. *J. Mach. Learn. Res.* **12** 1771–1812.

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

[9] DELLACHERIE, C., MARTINEZ, S. and SAN MARTIN, J. (2014). *Inverse M-Matrices and Ultrametric Matrices* **2118**. Springer, Berlin.

[10] DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

[11] DJOLONGA, J. and KRAUSE, A. (2015). Scalable variational inference in log-supermodular models. In *In International Conference on Machine Learning* (*ICML*).

[12] EGILMEZ, H. E., PAVEZ, E. and ORTEGA, A. (2016). Graph learning from data under structural and Laplacian constraints. Available at arXiv:1611.0518.

[13] FALLAT, S., LAURITZEN, S. L., SADEGHI, K., UHLER, C., WERMUTH, N. and ZWIERNIK, P. (2017). Total positivity in Markov structures. *Ann. Statist.* **45** 1152–1184.

[14] FELSENSTEIN, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25** 471–492.

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

[16] GOWER, J. C. and ROSS, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.* **18** 54–61. MR0242315

[17] GRANT, M. and BOYD, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. Available at http://cvxr.com/cvx.

[18] GROSS, E. and SULLIVANT, S. (2018). The maximum likelihood threshold of a graph. *Bernoulli* **24** 386–407.

[19] HØJSGAARD, S., EDWARDS, D. and LAURITZEN, S. (2012). *Graphical Models with R*. Springer, New York.

[20] JOHNSON, C. R. and SMITH, R. L. (1996). The completion problem for M-matrices and inverse M-matrices. *Linear Algebra Appl.* **241–243** 655–667.

[21] JOHNSON, C. R. and SMITH, R. L. (1999). Path product matrices. *Linear and Multilinear Algebra* **46** 177–191.

[22] JOHNSON, C. R. and SMITH, R. L. (2011). Inverse M-matrices, II. *Linear Algebra Appl.* **435** 953–983. MR2807211

[23] 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.

[24] KARLIN, S. and RINOTT, Y. (1981). Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *Ann. Statist.* **9** 1035–1049. MR0628759

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

[26] LAURITZEN, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

[27] LEDERMANN, W. (1940). On a problem concerning matrices with variable diagonal elements. *Proc. Roy. Soc. Edinburgh Sect. A* **60** 1–17. MR0001206

[28] LUO, Z. Q. and TSENG, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.* **72** 7–35.

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

[30] MALLE, B. F. and HOROWITZ, L. M. (1995). The puzzle of negative self-views: An exploration using the schema concept. *J. Pers. Soc. Psychol.* **68** 470.

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

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

[33] PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms* (*Atlanta, GA*, 1995) **9** 223–252. MR1611693

[34] SEMPLE, C. and STEEL, M. A. (2003). *Phylogenetics* **24**. Oxford Univ. Press, London.

[35] SHAPIRO, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Int. Stat. Rev.* **56** 49–62.

[36] SHIERS, N., ZWIERNIK, P., ASTON, J. and SMITH, J. Q. (2016). The correlation space of Gaussian latent tree models and model selection without fitting. *Biometrika* **103** 531–545.

[37] 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.

[38] SPEARMAN, C. (1928). The abilities of man. *Science* **68** 38.

[39] SPEED, T. P. and KIIVERI, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150. MR0829559

[40] UHLER, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.* **40** 238–261. MR3014306

[41] WERMUTH, N. and SCHEIDT, E. (1977). Algorithm AS 105: Fitting a covariance selection model to a matrix. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **26** 88–92.

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

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

C. UHLER
LABORATORY FOR INFORMATION AND DECISION SYSTEMS
AND INSTITUTE FOR DATA, SYSTEMS AND SOCIETY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: cuhler@mit.edu

P. ZWIERNIK
DEPARTMENT OF ECONOMICS AND BUSINESS
UNIVERSITAT POMPEU FABRA
RAMON TRIAS FARGAS, 25-27
08005 BARCELONA
SPAIN
E-MAIL: piotr.zwiernik@upf.edu