The Annals of Statistics
2019, Vol. 47, No. 3, 1536–1553
https://doi.org/10.1214/18-AOS1724

THE MAXIMUM LIKELIHOOD THRESHOLD OF A PATH DIAGRAM[1]

BY MATHIAS DRTON[*,†], CHRISTOPHER FOX[‡], ANDREAS KÄUFL[§] AND GUILLAUME POULIOT[‡,¶]

University of Washington[*], University of Copenhagen[†], University of Chicago[‡], University of Augsburg[§] and École Polytechnique de Montréal[¶]

Linear structural equation models postulate noisy linear relationships between variables of interest. Each model corresponds to a path diagram, which is a mixed graph with directed edges that encode the domains of the linear functions and bidirected edges that indicate possible correlations among noise terms. Using this graphical representation, we determine the maximum likelihood threshold, that is, the minimum sample size at which the likelihood function of a Gaussian structural equation model is almost surely bounded. Our result allows the model to have feedback loops and is based on decomposing the path diagram with respect to the connected components of its bidirected part. We also prove that if the sample size is below the threshold, then the likelihood function is almost surely unbounded. Our work clarifies, in particular, that standard likelihood inference is applicable to sparse high-dimensional models even if they feature feedback loops.

1. Introduction. Structural equation models are multivariate statistical models that treat each variable of interest as a function of the remaining variables and a random error term. Linear structural equation models require all these functions to be linear. Let $X = (X_1, \ldots, X_p)$ be the random vector holding the considered variables. Then X solves the equation system

$$X_i = \lambda_{0i} + \sum_{j \neq i} \lambda_{ij} X_j + \varepsilon_i, \qquad i = 1, \ldots, p, \tag{1.1}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$ is a given p-dimensional random error vector, and the λ_{0i} and λ_{ij} are unknown parameters. Let $\Lambda_0 = (\lambda_{01}, \ldots, \lambda_{0p})$ and form the matrix $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{p \times p}$ by setting the diagonal entries to zero. Following the frequently made Gaussian assumption, assume that ε is centered p-variate normal with covariance matrix $\Omega = (\omega_{ij})$. Writing I for the identity matrix, (1.1) yields that $X = (I - \Lambda)^{-T}\varepsilon$ is multivariate normal with covariance matrix

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}. \tag{1.2}$$

Here, and throughout the paper, the matrix $I - \Lambda$ is required to be invertible.

Received February 2018.
[1]Supported by NSF Grant DMS-1712535.
MSC2010 subject classifications. 62H12, 62J05.
Key words and phrases. Covariance matrix, graphical model, maximum likelihood, normal distribution, path diagram, structural equation model.

The Gaussian models just introduced have a long tradition [Wright (1921, 1934)] but remain an important tool for modern applications [e.g., Grace et al. (2016), Maathuis et al. (2010)]. Their popularity is driven by causal interpretability [Pearl (2009), Spirtes, Glymour and Scheines (2000)] as well as favorable statistical properties that facilitate analysis of highly multivariate data. In this paper, we focus on the fact that if the matrices Λ and Ω are suitably sparse, then maximum likelihood estimates in high-dimensional models may exist at small sample sizes. This enables, for instance, the use of likelihood in stepwise model selection. It can often be expected that Λ is sparse because each variable X_i depends on only a few of the other variables X_j, $j \neq i$. Similarly, the number of nonzero off-diagonal entries of Ω is small unless many pairs of error terms ε_i and ε_j are correlated through a latent common cause for X_i and X_j. We encode assumptions of sparsity in (Λ, Ω) in a graphical framework advocated by Wright (1921, 1934). Our terminology follows conventions from the book of Lauritzen (1996), the review of Drton (2016) and other related work such as Foygel, Draisma and Drton (2012) or Evans and Richardson (2016).

Background. A *mixed graph* is a triple $\mathcal{G} = (V, D, B)$ such that $D \subseteq V \times V$ and B is a set containing 2-element subsets of V. Throughout the paper, we take the vertex set to be $V = \{1, \ldots, p\}$ such that the nodes in V index the given random variables $X_1, \ldots, X_p$. The pairs $(i, j) \in D$ are *directed edges* that we denote as $i \to j$. Node j is the head of such an edge. We always assume that there are no self-loops, that is, $i \to i \notin D$ for all $i \in V$. The elements $\{i, j\} \in B$ are *bidirected edges* that have no orientation; we write such an edge as $i \leftrightarrow j$ or $j \leftrightarrow i$. Two nodes $i, j \in V$ are *adjacent* if $i \leftrightarrow j \in B$ or $i \to j \in D$ or $j \to i \in D$.

Let $\mathcal{G}' = (V', D', B')$ be another mixed graph. If $V' \subseteq V$, $D' \subseteq D$, and $B' \subseteq B$, then $\mathcal{G}'$ is a *subgraph* of $\mathcal{G}$, and $\mathcal{G}$ *contains* $\mathcal{G}'$. If $V' = \{i_0, i_1, \ldots, i_k\}$ for distinct $i_0, i_1, \ldots, i_k$ and there are $|D'| + |B'| = k$ edges such that any two consecutive nodes i_{h-1} and i_h are adjacent in $\mathcal{G}'$, then $\mathcal{G}'$ is a *path* from i_0 to i_k. It is a *directed path* if $i_{h-1} \to i_h$ for all h. Adding the edge $i_k \to i_0$ gives a *directed cycle*.

A mixed graph $\mathcal{G}$ is *connected* if it contains a path from any node i to any other node j. A *connected component* of $\mathcal{G}$ is an inclusion-maximal connected subgraph. In other words, a subgraph $\mathcal{G}'$ is a connected component of $\mathcal{G}$ if $\mathcal{G}'$ is connected and every subgraph of $\mathcal{G}$ that strictly contains $\mathcal{G}'$ fails to be connected. If $\mathcal{G}$ does not contain any directed cycles, then it is *acyclic*. If it has only directed edges ($B = \varnothing$), then $\mathcal{G}$ is a *digraph*. The graphical modeling literature refers to an *acyclic digraph* also as directed acyclic graph, abbreviated to DAG.

Now, let $\mathbb{R}^D$ be the space of real $p \times p$ matrices $\Lambda = (\lambda_{ij})$ with $\lambda_{ij} = 0$ whenever $i \to j \notin D$, and write $\mathbb{R}^D_{\mathrm{reg}}$ for the subset of matrices $\Lambda \in \mathbb{R}^D$ with $I - \Lambda$ invertible. Note that $\mathbb{R}^D = \mathbb{R}^D_{\mathrm{reg}}$ if and only if $\mathcal{G}$ is acyclic. Let $PD(B)$ be the cone of positive definite $p \times p$ matrices $\Omega = (\omega_{ij})$ with $\omega_{ij} = 0$ when $i \neq j$ and

$i \leftrightarrow j \notin B$. Then the linear structural equation model given by $\mathcal{G}$ is the set of multivariate normal distributions $\mathcal{N}(\mu, \Sigma)$ with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix Σ in

$$PD(\mathcal{G}) = \{(I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} : (\Lambda, \Omega) \in \mathbb{R}^D_{\text{reg}} \times PD(B)\}. \tag{1.3}$$

We remark that the graph $\mathcal{G}$ is also known as the *path diagram* of the model.

Maximum likelihood threshold. Suppose now that we have independent and identically distributed multivariate observations $X^{(1)}, \ldots, X^{(n)} \sim \mathcal{N}(\mu, \Sigma)$. Let

$$\bar{X}_n = \frac{1}{n} \sum_{s=1}^{n} X^{(s)} \quad \text{and} \quad S_n = \frac{1}{n} \sum_{s=1}^{n} (X^{(s)} - \bar{X}_n)(X^{(s)} - \bar{X}_n)^T \tag{1.4}$$

be the sample mean vector and sample covariance matrix. With an additive constant omitted and $n/2$ divided out, the log-likelihood function is

$$\ell(\mu, \Sigma | \bar{X}_n, S_n) = -\log\det(\Sigma) - \operatorname{trace}(\Sigma^{-1} S_n) - (\bar{X}_n - \mu)^T \Sigma^{-1} (\bar{X}_n - \mu).$$

The considered models have the mean vector unrestricted and the maximum likelihood estimator of μ is always $\bar{X}_n$. This yields the profile log-likelihood function

$$\ell(\Sigma | S_n) = -\log\det(\Sigma) - \operatorname{trace}(\Sigma^{-1} S_n). \tag{1.5}$$

Our interest is in determining, for a mixed graph $\mathcal{G} = (V, D, B)$, the minimum number N such that for a sample of size $n \geq N$ the log-likelihood function is almost surely bounded above on the set $\mathbb{R}^p \times PD(\mathcal{G})$. As usual, almost surely refers to probability one when $X^{(1)}, \ldots, X^{(n)}$ are an independent sample from a regular multivariate normal distribution, or equivalently, any other absolutely continuous distribution on $\mathbb{R}^p$. Let

$$\hat{\ell}(\mathcal{G} | S_n) = \sup\{\ell(\Sigma | S_n) : \Sigma \in PD(\mathcal{G})\}.$$

Adapting terminology from Gross and Sullivant (2018), the number we seek to derive is the *maximum likelihood threshold*

$$\operatorname{mlt}(\mathcal{G}) := \min\{N \in \mathbb{N} : \hat{\ell}(\mathcal{G} | S_n) < \infty \text{ a.s. } \forall n \geq N\}. \tag{1.6}$$

Here and throughout, a.s. abbreviates almost surely.

If we constrain the mean vector μ to be zero, then the relevant sample covariance matrix is

$$S_{0,n} = \frac{1}{n} \sum_{s=1}^{n} X^{(s)} (X^{(s)})^T. \tag{1.7}$$

By classical results [Anderson (2003), Chapter 7], $\operatorname{mlt}(\mathcal{G}) = \operatorname{mlt}_0(\mathcal{G}) + 1$, where

$$\operatorname{mlt}_0(\mathcal{G}) = \min\{N \in \mathbb{N} : \hat{\ell}(\mathcal{G} | S_{0,n}) < \infty \text{ a.s. } \forall n \geq N\} \tag{1.8}$$

is the maximum likelihood threshold for the model when the mean vector is taken to be zero. Our subsequent discussion will thus focus on the threshold $\operatorname{mlt}_0(\mathcal{G})$. We record three simple yet useful facts.

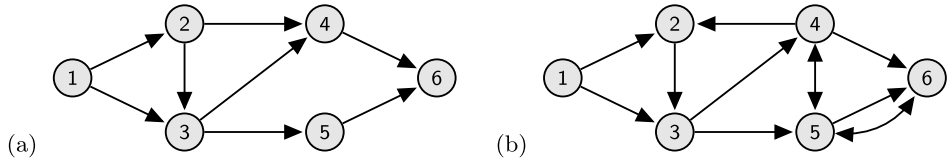

FIG. 1. (a) *An acyclic digraph with* $\mathrm{mlt}_0(\mathcal{G}) = 3$. (b) *A mixed graph with* $\mathrm{mlt}_0(\mathcal{G}) = 4$.

LEMMA 1. *Let* $\mathcal{G} = (V, D, B)$ *be a mixed graph. Then*

(a) $\mathrm{mlt}_0(\mathcal{G}) \leq p = |V|$.
(b) *If* $\mathcal{G}_1, \ldots, \mathcal{G}_k$ *are the connected components of* $\mathcal{G}$, *then*

$$\mathrm{mlt}_0(\mathcal{G}) = \max_{j=1,\ldots,k} \mathrm{mlt}_0(\mathcal{G}_j).$$

(c) *If* $\mathcal{H}$ *is a subgraph of* $\mathcal{G}$, *then* $\mathrm{mlt}_0(\mathcal{H}) \leq \mathrm{mlt}_0(\mathcal{G})$.

PROOF. (a) It is well known that $\ell(\cdot|S_{0,n})$ is bounded above on the entire cone of positive definite matrices if and only if $S_{0,n}$ is positive definite. Moreover, if S_n is positive, then $\Sigma = S_n$ is the unique maximizer [Anderson (2003), Lemma 3.2.2]. The matrix $S_{0,n}$ is positive definite a.s. if and only if $n \geq p$.

(b) The variables in the different connected components are independent. The likelihood function may be maximized separately for the different components.

(c) If $\mathcal{H}$ and $\mathcal{G}$ have the same vertex set, then $PD(\mathcal{H}) \subseteq PD(\mathcal{G})$ and, thus, $\hat{\ell}(\mathcal{H}|S_{0,n}) \leq \hat{\ell}(\mathcal{G}|S_{0,n})$. The case where $\mathcal{H}$ has fewer vertices can be addressed by adding isolated nodes and using the fact from (b). □

When $\mathcal{G}$ is connected, Lemma 1 yields only the trivial bound $\mathrm{mlt}_0(\mathcal{G}) \leq p$. However, $\mathrm{mlt}_0(\mathcal{G})$ may be far smaller than p when $\mathcal{G}$ is sparse, that is, has few edges. Indeed, in the well understood case of $\mathcal{G}$ being an acyclic digraph, maximum likelihood estimation reduces to solving one linear regression problem for each considered variable [Lauritzen (1996), page 154]. The predictors in the problem for variable j are the variables from the set of *parents* $\mathrm{pa}(j) = \{k \in V : k \to j \in D\}$. If the sample size exceeds the size of the largest parent set, then at least one degree of freedom remains for estimation of the error variance in each one of the p linear regression problems. We thus have the following well-known fact.

THEOREM 1. *Let* $\mathcal{G} = (V, D, \varnothing)$ *be an acyclic digraph. Then*

$$\mathrm{mlt}_0(\mathcal{G}) = 1 + \max_{j \in V} |\mathrm{pa}(j)|.$$

The quantity $|\mathrm{pa}(j)|$ in the theorem is also termed the in-degree of node j.

EXAMPLE 1. If $\mathcal{G}$ is the acyclic digraph from Figure 1(a), then the largest parent sets are of size two, for nodes $j \in \{3, 4, 6\}$. By Theorem 1, $\mathrm{mlt}_0(\mathcal{G}) = 3$.

Main result. In this paper, we determine $\mathrm{mlt}_0(\mathcal{G})$ for any mixed graph $\mathcal{G} = (V, D, B)$. For a set $A \subseteq V$, let $\mathrm{Pa}(A)$ be the union of A and the parents of its elements, so

$$\mathrm{Pa}(A) = A \cup \bigcup_{i \in A} \mathrm{pa}(i). \tag{1.9}$$

Then our main result can be stated as follows.

THEOREM 2. *Let $\mathcal{G} = (V, D, B)$ be a mixed graph, and let $C_1, \ldots, C_l$ be the vertex sets of the connected components of its bidirected part $\mathcal{G}_{\leftrightarrow} = (V, \varnothing, B)$. Then*

$$\mathrm{mlt}_0(\mathcal{G}) = \max_{j=1,\ldots,l} \bigl|\mathrm{Pa}(C_j)\bigr|.$$

Moreover, if $n < \mathrm{mlt}_0(\mathcal{G})$ then $\hat{\ell}(\mathcal{G}|S_{0,n}) = \infty$ a.s.

In the special case that $\mathcal{G}$ is an acyclic digraph, we have $B = \varnothing$ and Theorem 2 reduces to Theorem 1 because each connected component of $\mathcal{G}_{\leftrightarrow}$ has only a single node $j \in V$. Then $\mathrm{Pa}(\{j\}) = \mathrm{pa}(j) \cup \{j\}$ and $|\mathrm{Pa}(\{j\})| = 1 + |\mathrm{pa}(j)|$.

EXAMPLE 2. Let $\mathcal{G}$ be the graph in Figure 1(b). The parameters of the model given by $\mathcal{G}$ are identifiable in a generic or almost everywhere sense, as can be checked readily using the half-trek criterion [Barber, Drton and Weihs (2015), Foygel, Draisma and Drton (2012)]. Hence, $PD(\mathcal{G})$ is a 16-dimensional subset of the 21-dimensional cone of positive definite 6×6 matrices. By Theorem 2, $\mathrm{mlt}_0(\mathcal{G}) = 4$. Indeed, $\mathcal{G}_{\leftrightarrow}$ has four connected components with vertex sets $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$ and $C_4 = \{4, 5, 6\}$. Adding parents yields $\mathrm{Pa}(C_1) = \{1\}$, $\mathrm{Pa}(C_2) = \{1, 2, 4\}$, $\mathrm{Pa}(C_3) = \{1, 2, 3\}$ and $\mathrm{Pa}(C_4) = \{3, 4, 5, 6\}$.

REMARK 1. If the likelihood function associated with an acyclic digraph $\mathcal{G} = (V, D, \varnothing)$ is bounded, then it achieves its maximum. Hence, $n \geq \mathrm{mlt}_0(\mathcal{G})$ ensures that the maximum is a.s. achieved. We are not aware of any results in the literature that, for a more general class of graphs, would similarly guarantee achievement of the maximum. In fact, we believe that there are mixed graphs $\mathcal{G}$ such that even for sample size $n \geq \mathrm{mlt}_0(\mathcal{G})$ the probability of the likelihood function failing to achieve its maximum is not zero. This belief is based on the fact that the set $PD(\mathcal{G})$ is not generally closed. As a simple example, consider the graph $\mathcal{G}$ with edges $1 \to 2$, $2 \to 3$, and $2 \leftrightarrow 3$, for which $PD(\mathcal{G})$ comprises all positive definite 3×3 matrices $\Sigma = (\sigma_{ij})$ with $\sigma_{13} = 0$ whenever $\sigma_{12} = 0$.

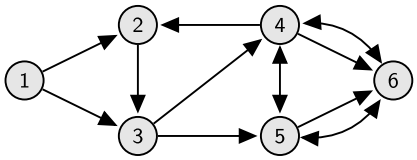

FIG. 2. *The graph $\mathcal{G}'$ when $\mathcal{G}$ is the mixed graph from Figure* 1(b). *Edge* $4 \leftrightarrow 6$ *has been added.*

Outline. In the remainder of the paper we first prove that $\mathrm{mlt}_0(\mathcal{G})$ is no larger than the value asserted in Theorem 2 (Section 2). Next, we derive $\mathrm{mlt}_0(\mathcal{G})$ for any bidirected graph $\mathcal{G}$ (Section 3). In Section 4, we use submodels given by bidirected graphs to show that the value from Theorem 2 is also a lower bound on $\mathrm{mlt}_0(\mathcal{G})$ for any (possibly cyclic) mixed graph, which then completes the proof of Theorem 2. A numerical experiment in Section 5 exemplifies that even a high-dimensional model is amenable to standard likelihood inference as long as its maximum likelihood threshold is small. The experiment suggests that likelihood inference allows one to perform model selection for high-dimensional but sparse cyclic models. In Section 6, we highlight interesting differences between the maximum likelihood threshold of Gaussian graphical models given by a directed versus an undirected cycle. The former model is nested in the latter and the two models have the same dimension, yet the thresholds are different.

2. Upper bound on the sample size threshold. We prove the upper bound that is part of Theorem 2.

THEOREM 3. *Let $\mathcal{G} = (V, D, B)$ be a mixed graph, and let $C_1, \ldots, C_l$ be the vertex sets of the connected components of its bidirected part $\mathcal{G}_{\leftrightarrow} = (V, \varnothing, B)$. Then*

$$\mathrm{mlt}_0(\mathcal{G}) \leq \max_{j=1,\ldots,l} \bigl|\mathrm{Pa}(C_j)\bigr|.$$

PROOF. Let $\mathcal{G}'$ be the supergraph of $\mathcal{G}$ obtained by adding bidirected edges between any two nodes that are in the same connected component of $\mathcal{G}_{\leftrightarrow} = (V, \varnothing, B)$ but that are not adjacent in $\mathcal{G}_{\leftrightarrow}$. Then $C_1, \ldots, C_l$ are still the vertex sets of the connected components of the bidirected part of $\mathcal{G}'$, and the sets $\mathrm{Pa}(C_j)$ are identical in $\mathcal{G}$ and $\mathcal{G}'$; see Figure 2 for an example. We emphasize that the bidirected part of $\mathcal{G}'$ is a disjoint union of complete subgraphs. The remainder of this proof shows the claimed bound for $\mathcal{G}'$. By Lemma 1(c), the bound then also holds for $\mathcal{G}$. To simplify notation, we assume that $\mathcal{G}$ itself has a bidirected part $\mathcal{G}_{\leftrightarrow}$ that is a disjoint union of complete graphs.

For $\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$, we have

$$\ell(\Sigma | S_{0,n}) = \log\bigl(\det(I - \Lambda)^2\bigr) - \log\det(\Omega) - \mathrm{trace}\bigl((I - \Lambda)\Omega^{-1}(I - \Lambda)^T S_{0,n}\bigr).$$

The set $PD(B)$ comprises all block-diagonal $p \times p$ matrices with l blocks determined by the connected components of $\mathcal{G}_{\leftrightarrow}$. Therefore, if $\Lambda = (\lambda_{jk}) \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega \in PD(B)$, we have

$$\begin{aligned}\ell(\Sigma|S_{0,n}) &= \log\bigl(\det(I-\Lambda)^2\bigr) \\ &\quad - \sum_{j=1}^{l}\bigl[\log\det(\Omega_{C_j,C_j}) \\ &\quad + \operatorname{trace}\bigl\{\Omega^{-1}_{C_j,C_j}\bigl((I-\Lambda)^T S_{0,n}(I-\Lambda)\bigr)_{C_j,C_j}\bigr\}\bigr].\end{aligned} \tag{2.1}$$

Let $X_1, \ldots, X_p$ be the columns of the data matrix

$$\mathbf{X} = \begin{pmatrix} X^{(1)} & \ldots & X^{(n)} \end{pmatrix}^T \in \mathbb{R}^{n\times p}.$$

Then $S_{0,n} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, and

$$\begin{aligned}\hat{\Omega}_{C_j,C_j} &= \bigl[(I-\Lambda)^T S_{0,n}(I-\Lambda)\bigr]_{C_j,C_j} \\ &= \frac{1}{n}\bigl[\mathbf{X}(I-\Lambda)_{V,C_j}\bigr]^T\bigl[\mathbf{X}(I-\Lambda)_{V,C_j}\bigr]\end{aligned}$$

is the sample covariance matrix of the vector of error terms

$$X_u - \sum_{k\in\mathrm{pa}(u)} \lambda_{ku} X_k,\ u \in C_j.$$

Fix $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$. Then, for any $j = 1, \ldots, l$, the function

$$\Omega_{C_j,C_j} \mapsto -\log\det(\Omega_{C_j,C_j}) - \operatorname{trace}\bigl\{\Omega^{-1}_{C_j,C_j}\bigl[(I-\Lambda)^T S_{0,n}(I-\Lambda)\bigr]_{C_j,C_j}\bigr\} \tag{2.2}$$

is bounded if and only if $\hat{\Omega}_{C_j,C_j}$ is positive definite. If it is bounded, then $\hat{\Omega}_{C_j,C_j}$ is the unique maximizer [Anderson (2003), Lemma 3.2.2]. We claim that if $n \geq |\mathrm{Pa}(C_j)|$, then $\hat{\Omega}_{C_j,C_j}$ is a.s. positive definite. Indeed, by the lemma in Okamoto (1973), all square submatrices of $\mathbf{X}$ are a.s. invertible. If $n \geq |\mathrm{Pa}(C_j)|$, this implies that the vectors X_k, $k \in \mathrm{Pa}(C_j)$, are a.s. linearly independent. The columns of $\mathbf{X}(I-\Lambda)_{V,C_j}$ are linear combinations of these vectors. Because $I-\Lambda$ is invertible, the submatrix $(I-\Lambda)_{V,C_j}$ has full column rank $|C_j|$. Therefore, $\mathbf{X}(I-\Lambda)_{V,C_j}$ a.s. has full column rank $|C_j|$, which implies positive definiteness of $\hat{\Omega}_{C_j,C_j}$.

Because a union of null sets is a null set, if $n \geq \max_{j=1,\ldots,l}|\mathrm{Pa}(C_j)|$, then a.s. all matrices $\hat{\Omega}_{C_j,C_j}$ for $j = 1, \ldots, l$ are simultaneously positive definite. We may thus proceed by substituting all $\hat{\Omega}_{C_j,C_j}$ into the log-likelihood function $\ell(\Sigma|S_{0,n})$ displayed in (2.1). The resulting profile log-likelihood function is

$$\begin{aligned}\ell(\Lambda|S_{0,n}) &= \log\bigl(\det(I-\Lambda)^2\bigr) - p \\ &\quad - \sum_{j=1}^{l}\log\det\bigl(\bigl[(I-\Lambda)^T S_{0,n}(I-\Lambda)\bigr]_{C_j,C_j}\bigr).\end{aligned} \tag{2.3}$$

In order to show that $\ell(\Lambda|S_{0,n})$ is a.s. bounded from above, we apply a block-version of the Hadamard inequality, which yields that

$$\log\bigl(\det(I-\Lambda)^2\bigr) \le \sum_{j=1}^{l} \log\det\bigl(\bigl[(I-\Lambda)^T(I-\Lambda)\bigr]_{C_j,C_j}\bigr); \tag{2.4}$$

recall that the sets C_j form a partition of $V=\{1,\dots,p\}$. Using (2.4) in (2.3), we see that up to a constant the exponential of $\ell(\Lambda|S_{0,n})$ is bounded above by the product of the terms

$$\begin{aligned}&\frac{\det([(I-\Lambda)^T(I-\Lambda)]_{C_j,C_j})}{\det([(I-\Lambda)^T S_{0,n}(I-\Lambda)]_{C_j,C_j})}\\ &\quad=\frac{\det((I-\Lambda^T)_{C_j,\mathrm{Pa}(C_j)}(I-\Lambda)_{\mathrm{Pa}(C_j),C_j})}{\det((I-\Lambda^T)_{C_j,\mathrm{Pa}(C_j)}(S_{0,n})_{\mathrm{Pa}(C_j),\mathrm{Pa}(C_j)}(I-\Lambda)_{\mathrm{Pa}(C_j),C_j})}\end{aligned} \tag{2.5}$$

for $j=1,\dots,l$. Let $\lambda_j(S_{0,n})$ be the minimum eigenvalue of the $\mathrm{Pa}(C_j)\times\mathrm{Pa}(C_j)$ submatrix of $S_{0,n}$. This submatrix is the sample covariance matrix of the variables indexed by $\mathrm{Pa}(C_j)$. Therefore, if $n\ge|\mathrm{Pa}(C_j)|$, then $\lambda_j(S_{0,n})$ is a.s. positive. Now, $(S_{0,n})_{\mathrm{Pa}(C_j),\mathrm{Pa}(C_j)}\succeq\lambda_j(S_{0,n})I$ in the positive semidefinite ordering. Using Observation 7.2.2 and Corollary 7.7.4(b) in Horn and Johnson (1990), we obtain that the ratio in (2.5) is a.s. bounded above by $\lambda_j(S_{0,n})^{-|C_j|}<\infty$. $\square$

3. Bidirected graphs. Consider a bidirected graph $\mathcal{G}=(V,\varnothing,B)$. Then $PD(\mathcal{G})=PD(B)$ is a set of sparse positive definite matrices. We prove that the bound from Lemma 1(a) is an equality when the bidirected graph $\mathcal{G}$ is connected.

THEOREM 4. *If $\mathcal{G}=(V,\varnothing,B)$ is connected, then $\mathrm{mlt}_0(\mathcal{G})=p$. Moreover, if $n<\mathrm{mlt}_0(\mathcal{G})$ then $\hat{\ell}(\mathcal{G}|S_{0,n})=\infty$ a.s.*

The proof of the theorem makes use of two lemmas. We derive those first.

LEMMA 2. *If $n<p$, then the kernel of $S_{0,n}$ a.s. contains a vector $q\in\mathbb{R}^p$ with all coordinates nonzero.*

PROOF. The matrix $S_{0,n}$ has the same kernel as

$$\mathbf{X}=\begin{pmatrix}X^{(1)} & \dots & X^{(n)}\end{pmatrix}^T\in\mathbb{R}^{n\times p}.$$

Partition the matrix as $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2)$, where the square submatrix $\mathbf{X}_1$ contains the first n columns. The determinant being a polynomial, the lemma in Okamoto (1973) yields that $\mathbf{X}_1$ is a.s. invertible.

We claim that for all $j\le n$, the kernel of $\mathbf{X}$ almost surely contains a vector q with $q_{n+1}=\cdots=q_p=1$ and $q_j\ne0$. Without loss of generality, it suffices

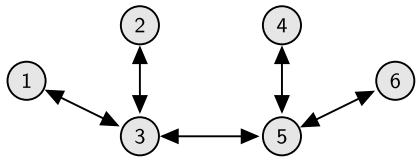

FIG. 3. *A bidirected graph labeled in such a way that for any j the nodes $i \geq j$ induce a connected subgraph.*

to treat the case of $j = 1$. By the above discussion, we may assume that $\mathbf{X}_1$ is invertible. Then a partitioned vector $(u, v) \in \mathbb{R}^p$ is in the kernel of $\mathbf{X}$ if and only if $u = -\mathbf{X}_1^{-1}\mathbf{X}_2 v$. Let $e = (1, \ldots, 1)^T \in \mathbb{R}^{p-n}$. The claim is true if and only if the vector $u = -\mathbf{X}_1^{-1}\mathbf{X}_2 e$ has first entry $u_1 \neq 0$. Multiplying u_1 with $\det(\mathbf{X}_1)$ gives a polynomial $f(\mathbf{X})$ such that $u_1 = 0$ only if $f(\mathbf{X}) = 0$. The lemma in Okamoto (1973) yields the claim if we can argue that the product $f(\mathbf{X})\det(\mathbf{X}_1)$ is not the zero polynomial. To this end, it is enough to exhibit one matrix $\mathbf{X}$ such that $u_1 \neq 0$ and $\det(\mathbf{X}_1) \neq 0$. Take $\mathbf{X}_1 = I$ and let $\mathbf{X}_2$ have a single nonzero entry $\mathbf{X}_{1,n+1} = -1$. Then $u = (1, 0, \ldots, 0)^T$.

Because a union of null sets is a null set, the kernel of $\mathbf{X}$ almost surely contains a vector q with $q_{n+1} = \cdots = q_p = 1$ and $q_j \neq 0$ for all $j \leq n$. $\square$

LEMMA 3. *Let q be any vector with all entries nonzero. There exists a matrix $\Sigma \in PD(\mathcal{G})$, such that the vector Σq has precisely one nonzero entry.*

PROOF. For a subset of nodes $A \subset V$, let $\mathcal{G}_A = (A, \varnothing, B_A)$ be the subgraph of $\mathcal{G}$ induced by A, that is, $B_A = B \cap (A \times A)$. Since $\mathcal{G}$ is connected, we may assume that the vertex set $V = \{1, \ldots, p\}$ has been relabeled such that the induced subgraph $\mathcal{G}_{\{i+1,\ldots,p\}}$ is connected for all $i = 1, \ldots, p-1$ [Diestel (2010), Proposition 1.4.1]. Figure 3 shows an example of a bidirected graph that is labeled in this way.

We now show how to construct $\Sigma = (\sigma_{kl}) \in PD(\mathcal{G})$ such that $(\Sigma q)_j = 0$ for all $j < p$. Since $q \neq 0$ and Σ will be positive definite, we then have $(\Sigma q)_p \neq 0$. As Σ must be symmetric, we only have to specify the entries $\bar{\sigma}_{kl}$ with $k \leq l$.

We construct Σ one row (and by symmetry, column) at a time according to the following iterative procedure. At stage $i = 1, \ldots, p$, the first $i-1$ rows and columns have been specified; none when $i = 1$. Let $\Sigma_{[i],[i]} = (\sigma_{kl})_{k,l \leq i}$ be the ith leading principal submatrix. We set σ_{ii} to be the smallest natural number with the property that $\det(\Sigma_{[i],[i]}) > 0$; that such a choice is possible is clear from a Laplace expansion of the determinant. For $i = 1$, we get $\sigma_{ii} = 1$. Next, as long as $i < p$, we choose $i^* \in \{i+1, \ldots, p\}$ such that $i \leftrightarrow i^* \in B$, which is possible because $\mathcal{G}_{\{i,\ldots,p\}}$ is connected. For all $k \geq i+1$ and $k \neq i^*$, we set $\sigma_{ik} = 0$ if $i \leftrightarrow k \notin B$ and $\sigma_{ik} = 1$ if $i \leftrightarrow k \in B$. We then complete the ith row and column by

setting

$$\sigma_{ii^*} = -\sum_{l \in V \setminus \{i^*\}} \sigma_{il} q_l / q_{i^*};$$

the division by q_{i^*} is well defined as all entries of q are nonzero.

By construction, the matrix Σ is positive definite as all leading principal minors are positive. Moreover, $\Sigma_{ij} = 0$ whenever $i \neq j$ and $i \leftrightarrow j \notin B$. It follows that $\Sigma \in PD(B) = PD(\mathcal{G})$. Finally, for all $i \leq p-1$,

$$(\Sigma q)_i = \sigma_{ii^*} q_{i^*} + \sum_{l \neq i^*} \sigma_{il} q_l = -\Big(\sum_{l \neq i^*} \sigma_{il} \frac{q_l}{q_{i^*}}\Big) q_{i^*} + \sum_{l \neq i^*} \sigma_{kl} q_l = 0. \qquad \square$$

PROOF OF THEOREM 4. By Lemma 1(a), we have $\mathrm{mlt}_0(\mathcal{G}) \leq p$. Hence, we need to show that the likelihood function is a.s. unbounded if $n < p$.

Assume that $n < p$. By Lemma 2, the kernel of the sample covariance matrix $S_{0,n}$ a.s. contains a vector q with all entries nonzero. By Lemma 3, we may choose a matrix Σ such that Σq has one nonzero entry. Without loss of generality, we assume the vertex set to be labeled such that $\Sigma q = c e_p$, where $c \in \mathbb{R} \setminus \{0\}$ and $e_p = (0, \ldots, 0, 1)^T \in \mathbb{R}^p$. Based on these choices, we will define a sequence of covariance matrices $\{\Sigma_t\}_{t=1}^{\infty}$ in $PD(\mathcal{G})$, with the property that $\lim_{t \to \infty} \ell(\Sigma_t | S_{0,n}) = \infty$. This then implies that the likelihood function is a.s. unbounded.

For $t \geq 0$, define

$$\Sigma_t := \Sigma - \frac{1}{\frac{1}{t} + q^T \Sigma q} \Sigma q q^T \Sigma. \tag{3.1}$$

Since $\Sigma q = c e_p$, the matrix $(\Sigma q)(\Sigma q)^T = c^2 e_p e_p^T$ is zero with the exception of the (p, p) entry that equals $c^2 > 0$. Hence, Σ_t has zeros in the same entries as $\Sigma \in PD(\mathcal{G})$ does. Let $K = (\Sigma)^{-1}$. By the Woodbury matrix identity [Woodbury (1950)],

$$K_t := (\Sigma_t)^{-1} = K + t q q^T.$$

For all $t \geq 0$, the matrix K_t is positive definite because K is positive definite and qq^T positive semidefinite. Thus, Σ_t is positive definite for all $t \geq 0$ as well. We conclude that $\Sigma_t \in PD(\mathcal{G})$ for all $t \geq 0$.

Inserting Σ_t into the log-likelihood function from (1.5), we have

$$\begin{aligned} \ell(\Sigma_t | S_{0,n}) &= \log\det(K_t) - \mathrm{trace}(K_t S_{0,n}) \\ &= \log\det\big(K + t q q^T\big) - \mathrm{trace}(K S_{0,n}) - t q^T S_{0,n} q \\ &= \log\det\big(K + t q q^T\big) - \mathrm{trace}(K S_{0,n}) \end{aligned}$$

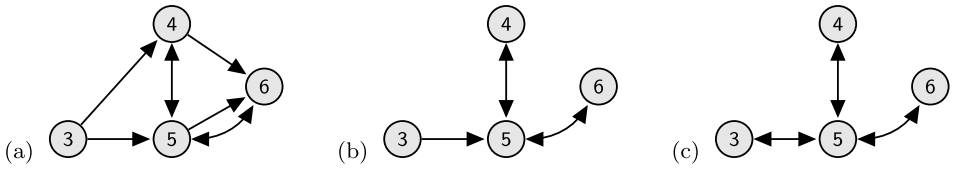

FIG. 4. *In reference to the graph from Figure* 1(b), *the panels show*: (a) *the connected component* $\mathcal{G}_4$ *with vertex set* $C_4 = \{4, 5, 6\}$, (b) *a choice for* $\mathcal{H}_4 \subset \mathcal{G}_4$ *and* (c) *the bidirected graph* $\mathcal{H}_4^{\leftrightarrow}$.

because q is in the kernel of $S_{0,n}$. By the matrix determinant lemma,

$$\det(K + tqq^T) = (1 + tq^T \Sigma q) \det(\Sigma),$$

which converges to infinity as $t \to \infty$ because $\det(\Sigma) > 0$ and $q^T \Sigma q > 0$ by positive definiteness of Σ. $\square$

4. Lower bound from submodels. We return to the case where $\mathcal{G} = (V, D, B)$ is an arbitrary, possibly cyclic, mixed graph. The following result uses the characterization of the maximum likelihood threshold for bidirected graphs to yield a lower bound on $\mathrm{mlt}_0(\mathcal{G})$.

THEOREM 5. *Let* $\mathcal{G} = (V, D, B)$ *be a mixed graph, and let* $C_1, \ldots, C_l$ *be the vertex sets of the connected components of its bidirected part* $\mathcal{G}_{\leftrightarrow} = (V, \varnothing, B)$. *Then*

$$\mathrm{mlt}_0(\mathcal{G}) \geq \max_{j=1,\ldots,l} |\mathrm{Pa}(C_j)|.$$

Moreover, if $n < \mathrm{mlt}_0(\mathcal{G})$ *then* $\hat{\ell}(\mathcal{G}|S_{0,n}) = \infty$ *a.s.*

PROOF. For $j = 1, \ldots, l$, let $B_j = B \cap (C_j \times C_j)$ and $D_j = D \cap (\mathrm{Pa}(C_j) \times C_j)$. In other words, B_j is the set of bidirected edges between nodes in C_j, while D_j is the set of directed edges with head in C_j. The sets B_j and D_j partition B and D, respectively. The graphs $\mathcal{G}_j = (\mathrm{Pa}(C_j), D_j, B_j)$ thus form a decomposition of $\mathcal{G}$. Because each graph $\mathcal{G}_j$ is a subgraph of $\mathcal{G}$, Lemma 1(c) yields that

$$\mathrm{mlt}_0(\mathcal{G}) \geq \max_{j=1,\ldots,l} \mathrm{mlt}_0(\mathcal{G}_j).$$

Next, for each j, choose a subgraph $\mathcal{H}_j$ of $\mathcal{G}_j$ by taking the bidirected part of $\mathcal{G}_j$ and adding for each node in $\mathrm{Pa}(C_j) \setminus C_j$ precisely one of its outgoing directed edges. Then let $\mathcal{H}_j^{\leftrightarrow}$ be the bidirected graph obtained by converting the directed edges of $\mathcal{H}_j$ into bidirected edges. An example is shown in Figure 4. Since in $\mathcal{H}_j$ each node $i \in \mathrm{Pa}(C_j) \setminus C_j$ is the parent of precisely one node in C_j, it follows from Theorem 5 in Drton and Richardson (2008) that $\mathcal{H}_j$ and $\mathcal{H}_j^{\leftrightarrow}$ define the same set of covariance matrices. Consequently,

$$PD(\mathcal{H}_j^{\leftrightarrow}) = PD(\mathcal{H}_j) \subseteq PD(\mathcal{G}_j).$$

Now use Lemma 1(c) and apply Theorem 4 to the connected bidirected graph $\mathcal{H}_j^{\leftrightarrow}$ to conclude that

$$\mathrm{mlt}_0(\mathcal{G}_j) \geq \mathrm{mlt}_0(\mathcal{H}_j) = \mathrm{mlt}_0\big(\mathcal{H}_j^{\leftrightarrow}\big) = \big|\mathrm{Pa}(C_j)\big|. \qquad \square$$

5. Numerical experiment. A model with low maximum likelihood threshold is amenable to standard likelihood inference even when the modeled observations are high dimensional and the sample size is rather small. We demonstrate this for a structural equation model associated with a directed graph and allowing for cycles. Specifically, we consider a graph $\mathcal{G}_p = (V_p, E_p)$ with an even number p of nodes. As previously, we enumerate the vertex set as $V_p = \{1, \ldots, p\}$. Let $p' = p/2$, and define the edge set as $E_p = E_p^{(1)} \cup E_p^{(2)} \cup E_p^{(3)}$, where

$$\begin{aligned}
E_p^{(1)} &= \big\{i \to i+1 : i = 1, \ldots, p'-1\big\} \cup \big\{p' \to 1\big\},\\
E_p^{(2)} &= \big\{i+3 \to i : i = 1, \ldots, p'-3\big\} \cup \big\{1 \to p'-2\big\}\\
&\qquad \cup \big\{2 \to p'-1\big\} \cup \big\{3 \to p'\big\},\\
E_p^{(3)} &= \big\{p'+i \to i : i = 1, \ldots, p'\big\}.
\end{aligned}$$

The first set of edges defines a directed cycle of length p', and the second set of edges gives many shorter cycles of length 4. The third set of edges attaches, in bipartite fashion, additional nodes that play the role of covariates; one covariate for each node in the long cycle. Figure 5 illustrates this with a picture of $\mathcal{G}_{40}$.

As a statistical problem we consider testing absence of the edge $1 \to 2$ from the graph $\mathcal{G}_{100}$. In other words, we test the hypothesis $H_0 : \lambda_{12} = 0$ in the model given by $\mathcal{G}_{100}$. The parametrization for $\mathcal{G}_{100}$ is generically one-to-one as can be confirmed, for instance, using the half-trek criterion [Barber, Drton and Weihs (2015), Foygel, Draisma and Drton (2012)]. Assuming zero means for the $p = 100$ dimensional observation vector, the model corresponds to a $p + 3p/2 = 250$ dimensional set of covariance matrices. We test H_0 using the likelihood ratio test for three rather small sample sizes, namely, $n = 15$, 20 and 25. Our main result guarantees that the test is well defined as the log-likelihood function for $\mathcal{G}_{100}$ a.s. admits a finite supremum at these sample sizes. The optimization needed to compute the likelihood ratio statistics is performed using the algorithm of Drton, Fox and Wang (2018).

For each sample size, we use 200 Monte Carlo simulations to approximate the size of the test as well as its power at nonzero values of λ_{12}. Specifically, we consider the setting where λ_{12} ranges through $[-1, 1]$, and all other edge coefficients are set to $1/3$. We consider nominal significance level 0.05 and calibrate the likelihood ratio test using a chi-square distribution with 1 degree of freedom. A chi-square limiting distribution cannot always be expected [Drton (2009)], but is valid at the considered identifiable parameter. The power functions are plotted in Figure 6. The asymptotically calibrated test clearly exhibits good power at stronger

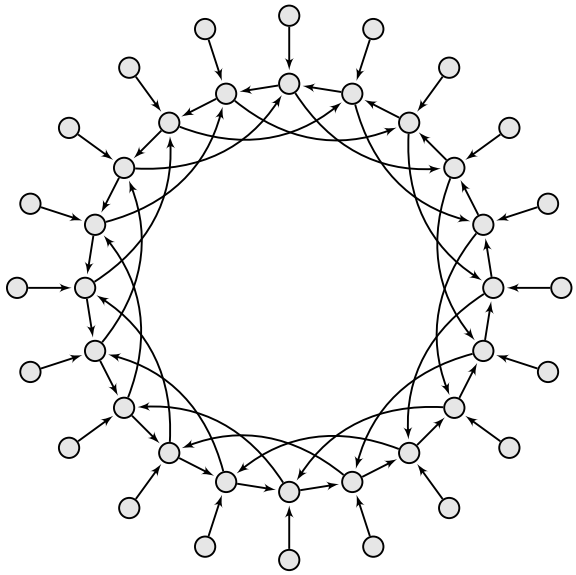

FIG. 5. *A directed graph with cycles and maximum in-degree* 3.

signals and is seen to be only slightly liberal. This suggests that likelihood inference allows one to perform model selection in high-dimensional but sparse cyclic models.

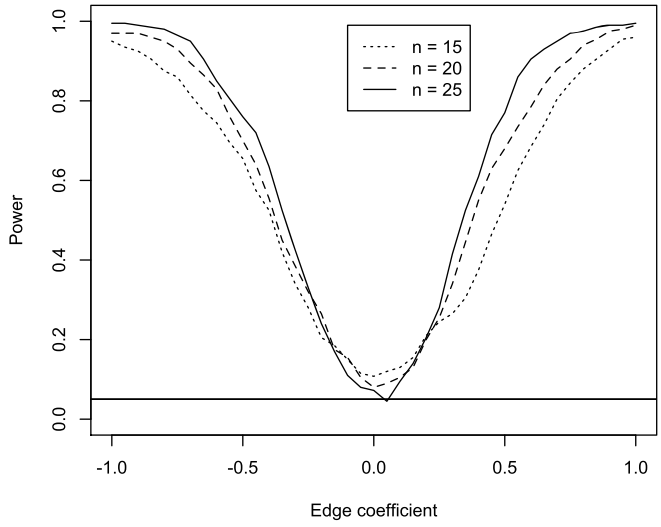

FIG. 6. *Monte Carlo approximation to power function of a likelihood ratio test.*

6. Connections to undirected graphical models. The structural equation models we considered are closely related to the Gaussian graphical models given by undirected graphs [Lauritzen (1996)]. The latter are dual to the models given by bidirected graphs in the sense that it is not the covariance matrix but its inverse that is supported over the graph [Kauermann (1996)]. To be more precise, let $\bar{\mathcal{G}} = (V, E)$ be an undirected graph, whose edges we take to be unordered pairs $\{i, j\}$ comprised of two distinct nodes $i, j \in V = \{1, \ldots, p\}$. Let $PD(\bar{\mathcal{G}})$ be the cone of positive definite $p \times p$ matrices $K = (\kappa_{ij})$ with $\kappa_{ij} = 0$ whenever $i \neq j$ and $\{i, j\} \notin E$. The Gaussian graphical model given by $\bar{\mathcal{G}}$ is the set of multivariate normal distributions $\mathcal{N}(\mu, \Sigma)$ with arbitrary mean vector $\mu \in \mathbb{R}^p$ and a covariance matrix that is constrained to have $\Sigma^{-1} \in PD(\bar{\mathcal{G}})$.

Suppose $\mathcal{G} = (V, D, \varnothing)$ is an acyclic digraph that is perfect, that is, $i, j \in \mathrm{pa}(k)$ implies that i and j are adjacent. Then $PD(\mathcal{G})$ is equal to the set of covariance matrices of the Gaussian graphical model given by the skeleton of $\mathcal{G}$ [see, e.g., Andersson, Madigan and Perlman (1997), Corollaries 4.1, 4.3]. The skeleton is the undirected graph $\mathcal{G}_- = (V, E)$ with $\{i, j\} \in E$ whenever i and j are adjacent in D. When $\mathcal{G}$ is perfect then $\mathcal{G}_-$ is chordal. Theorem 1 implies that the maximum likelihood threshold of the Gaussian graphical model of a chordal graph is the maximum clique size; see Grone et al. (1984), Theorem 7 or Buhl (1993), Theorem 3.2.

The maximum likelihood threshold of graphical models given by nonchordal graphs is more subtle to derive. Many interesting results exist, but the threshold has not yet been determined in generality [Buhl (1993), Gross and Sullivant (2018), Uhler (2012)]. Moreover, it has been shown for a sample size below the threshold that the likelihood may be bounded with positive probability. In the remainder of this section we focus on chordless cycles, which were the first known examples of this phenomenon. We note that in the literature the maximum likelihood threshold for Gaussian graphical models is typically introduced as the minimum sample size at which the likelihood function admits a maximizer. The maximizer is then unique by strict convexity of the log-likelihood function as a function of the inverse covariance matrix. By the duality theory in Dahl, Vandenberghe and Roychowdhury (2008), if the likelihood function of a Gaussian graphical model does not achieve its maximum, then it is unbounded; see also Theorem 9.5 in Barndorff-Nielsen (1978).

EXAMPLE 3. Let $\bar{\mathcal{C}}_p$ be the undirected chordless cycle with vertex set $V = \{1, \ldots, p\}$ and edge set $E = \{\{1, 2\}, \{2, 3\}, \ldots, \{p-1, p\}, \{1, p\}\}$ for $p \geq 3$. Assuming the mean vector μ to be zero, the Gaussian graphical model given by $\bar{\mathcal{C}}_p$ has maximum likelihood threshold $\mathrm{mlt}_0(\bar{\mathcal{C}}_p) = 3$. However, if $n = 2$, then the likelihood function of the model with zero means is bounded, and achieves its maximum, with positive probability [Buhl (1993), Theorem 4.1].

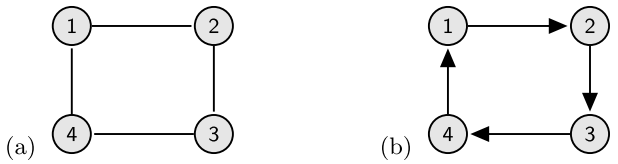

FIG. 7. (a) *The undirected cycle* $\bar{\mathcal{C}}_4$. (b) *The directed cycle* $\mathcal{C}_4$.

Let $PD(\bar{\mathcal{C}}_p)^{-1}$ be the set of matrices with an inverse in $PD(\bar{\mathcal{C}}_p)$. In other words, $PD(\bar{\mathcal{C}}_p)^{-1}$ is the set of covariance matrices of the graphical model for $\bar{\mathcal{C}}_p$. If an acyclic digraph $\mathcal{G} = (V, D, \varnothing)$ satisfies $PD(\mathcal{G}) \subseteq PD(\bar{\mathcal{C}}_p)^{-1}$, then $PD(\mathcal{G})$ has smaller dimension than $PD(\bar{\mathcal{C}}_p)^{-1}$. If $PD(\mathcal{G}) \supseteq PD(\bar{\mathcal{C}}_p)^{-1}$, then the dimension of $PD(\mathcal{G})$ is larger. However, a subset of the same dimension is found when considering digraphs with cycles. Specifically, take $\mathcal{C}_p = (V, D, \varnothing)$ to be the digraph with vertex set $V = \{1, \ldots, p\}$ and edge set $D = \{1 \to 2, 2 \to 3, \ldots, p-1 \to p, p \to 1\}$. Then $PD(\mathcal{C}_p) \subseteq PD(\bar{\mathcal{C}}_p)^{-1}$. Indeed, if $\Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}$ for $\Lambda \in \mathbb{R}^D_{\text{reg}}$ and $\Omega \in PD(B)$, then $\Sigma^{-1} = (I - \Lambda)\Omega^{-1}(I - \Lambda)^T$ has entries

$$\Sigma^{-1}_{ij} = \begin{cases} \dfrac{1}{\omega_{ii}} + \dfrac{\lambda^2_{i,i+1}}{\omega_{i+1,i+1}} & \text{if } i = j, \\ -\dfrac{\lambda_{i,i+1}}{\omega_{i+1,i+1}} & \text{if } j \in \{i-1, i+1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Here, we identify $p + 1 \equiv 1$. Recall that for a digraph $\Omega = (\omega_{ij})$ is diagonal. The zeros in (6.1) now confirm that $PD(\mathcal{C}_p) \subseteq PD(\bar{\mathcal{C}}_p)^{-1}$. Moreover, $PD(\mathcal{C}_p)$ is a full-dimensional subset as both $PD(\mathcal{C}_p)$ and $PD(\bar{\mathcal{C}}_p)$ clearly have dimension $2p$.

EXAMPLE 4. The graphs $\bar{\mathcal{C}}_4$ and $\mathcal{C}_4$ are depicted in Figure 7. A matrix in $PD(\mathcal{C}_4)$ is parameterized as

$$\begin{pmatrix} \dfrac{1}{\omega_{11}} + \dfrac{\lambda^2_{12}}{\omega_{22}} & -\dfrac{\lambda_{12}}{\omega_{22}} & 0 & -\dfrac{\lambda_{41}}{\omega_{11}} \\ -\dfrac{\lambda_{12}}{\omega_{22}} & \dfrac{1}{\omega_{22}} + \dfrac{\lambda^2_{23}}{\omega_{33}} & -\dfrac{\lambda_{23}}{\omega_{33}} & 0 \\ 0 & -\dfrac{\lambda_{23}}{\omega_{33}} & \dfrac{1}{\omega_{33}} + \dfrac{\lambda^2_{34}}{\omega_{44}} & -\dfrac{\lambda_{34}}{\omega_{44}} \\ -\dfrac{\lambda_{41}}{\omega_{11}} & 0 & -\dfrac{\lambda_{34}}{\omega_{44}} & \dfrac{1}{\omega_{44}} + \dfrac{\lambda^2_{41}}{\omega_{11}} \end{pmatrix}. \tag{6.2}$$

By Theorem 4.1 in Buhl (1993), $\text{mlt}_0(\bar{\mathcal{C}}_p) = 3$ for all $p \geq 3$. In contrast, our new Theorem 2 implies that $\text{mlt}_0(\mathcal{C}_p) = 2$ for all $p \geq 3$. Consequently, it must hold that $PD(\mathcal{C}_p) \subsetneq PD(\bar{\mathcal{C}}_p)^{-1}$. Indeed, the set $PD(\mathcal{C}_p)$ comprises matrices that

satisfy an additional inequality. Applying a trick used by Drton and Yu (2010) in a different context, observe that negating the entry λ_{12} changes only the entries Σ_{12}^{-1} and Σ_{21}^{-1}, which are negated. All other entries of Σ^{-1} are preserved under the sign change of λ_{12}. The inequality is obtained by noting that not every positive definite matrix in $PD(\mathcal{C}_p)$ remains positive definite after negation of a single off-diagonal entry [Drton and Yu (2010), Example 5.2]. We conclude that if a sample of size 2 has the likelihood function given by $\bar{\mathcal{C}}_p$ unbounded, then the divergence occurs only along sequences of matrices that do not represent a system with a feedback cycle as in $\mathcal{C}_p$.

7. Discussion. Our main result, Theorem 2, determines the maximum likelihood threshold of any linear structural equation model. This threshold is the smallest integer N such that the Gaussian likelihood function is a.s. bounded for all samples of size at least N. According to our result, the maximum likelihood threshold of models with feedback loops is surprisingly low. Indeed, the maximum likelihood threshold of any digraph, acyclic or not, is equal to the maximum in-degree plus one. In contrast, bidirected edges, which represent the effects of unmeasured confounders, can result in a large maximum likelihood threshold by merely forming long paths. If $\mathcal{G}$ is a bidirected spanning tree, then there are only $p-1$ edges yet $\mathrm{mlt}_0(\mathcal{G}) = p$, which is the largest possible value by Lemma 1(a).

When the structural equation model is given by an acyclic digraph, boundedness of the likelihood function implies that the maximum is achieved. As we emphasized in Remark 1, the question of when the maximum is a.s. achieved is still poorly understood for general mixed graphs and constitutes an important topic for future work.

Acknowledgment. We would like to thank Caroline Uhler for helpful discussions.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

BARBER, R. F., DRTON, M. and WEIHS, L. (2015). SEMID: Identifiability of linear structural equation models. R package version 0.2.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333

BUHL, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Stat.* **20** 263–270. MR1241392

DAHL, J., VANDENBERGHE, L. and ROYCHOWDHURY, V. (2008). Covariance selection for nonchordal graphs via chordal embedding. *Optim. Methods Softw.* **23** 501–520. MR2440363

DIESTEL, R. (2010). *Graph Theory*, 4th ed. *Graduate Texts in Mathematics* **173**. Springer, Heidelberg. MR2744811

DRTON, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658

DRTON, M. (2016). Algebraic problems in structural equation modeling. Available at arXiv:1612.05994.

DRTON, M., FOX, C. and WANG, Y. S. (2018). Computation of maximum likelihood estimates in cyclic structural equation models. *Ann. Statist.* To appear. Available at arXiv:1610.03434.

DRTON, M. and RICHARDSON, T. S. (2008). Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Mach. Learn. Res.* **9** 893–914. MR2417257

DRTON, M. and YU, J. (2010). On a parametrization of positive semidefinite matrices with zeros. *SIAM J. Matrix Anal. Appl.* **31** 2665–2680. MR2740626

EVANS, R. J. and RICHARDSON, T. S. (2016). Smooth, identifiable supermodels of discrete DAG models with latent variables. Available at arXiv:1511.06813.

FOYGEL, R., DRAISMA, J. and DRTON, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.* **40** 1682–1713. MR3015040

GRACE, J. B., ANDERSON, T. M., SEABLOOM, E. W., BORER, E. T., ADLER, P. B., HARPOLE, W. S., HAUTIER, Y., HILLEBRAND, H., LIND, E. M., PÄRTEL, M., BAKKER, J. D., BUCKLEY, Y. M., CRAWLEY, M. J., DAMSCHEN, E. I., DAVIES, K. F., FAY, P. A., FIRN, J., GRUNER, D. S., HECTOR, A., KNOPS, J. M. H., MACDOUGALL, A. S., MELBOURNE, B. A., MORGAN, J. W., ORROCK, J. L., PROBER, S. M. and SMITH, M. D. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature* **529** 390–393.

GRONE, R., JOHNSON, C. R., DE SÁ, E. M. and WOLKOWICZ, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58** 109–124. MR0739282

GROSS, E. and SULLIVANT, S. (2018). The maximum likelihood threshold of a graph. *Bernoulli* **24** 386–407. MR3706762

HORN, R. A. and JOHNSON, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

KAUERMANN, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Stat.* **23** 105–116. MR1380485

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

MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

OKAMOTO, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.* **1** 763–765. MR0331643

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

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

WOODBURY, M. A. (1950). *Inverting Modified Matrices*. *Statistical Research Group*, *Memo. Rep.* **42**. Princeton Univ., Princeton, NJ. MR0038136

WRIGHT, S. (1921). Correlation and causation. *J. Agricultural Research* **20** 557–585.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Stat.* **5** 161–215.

M. DRTON
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195-4322
USA
AND
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITETSPARKEN 5
2100 COPENHAGEN Ø
DENMARK
E-MAIL: md5@uw.edu

A. KÄUFL
INSTITUTE FOR MATHEMATICS
UNIVERSITY OF AUGSBURG
86159 AUGSBURG
GERMANY
E-MAIL: andreas.kaeufl@googlemail.com

C. FOX
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: chrisfox@uchicago.edu

G. POULIOT
HARRIS SCHOOL OF PUBLIC POLICY
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
AND
DEPARTMENT OF MATHEMATICS
AND INDUSTRIAL ENGINEERING
ÉCOLE POLYTECHNIQUE DE MONTRÉAL
MONTRÉAL, QC H3T 1J4
CANADA
E-MAIL: guillaume.pouliot@gmail.com