The Annals of Statistics
2019, Vol. 47, No. 2, 663–690
https://doi.org/10.1214/17-AOS1602

COMPUTATION OF MAXIMUM LIKELIHOOD ESTIMATES IN CYCLIC STRUCTURAL EQUATION MODELS

BY MATHIAS DRTON*, CHRISTOPHER FOX† AND Y. SAMUEL WANG*

*University of Washington** *and University of Chicago*†

Software for computation of maximum likelihood estimates in linear structural equation models typically employs general techniques from nonlinear optimization, such as quasi-Newton methods. In practice, careful tuning of initial values is often required to avoid convergence issues. As an alternative approach, we propose a block-coordinate descent method that cycles through the considered variables, updating only the parameters related to a given variable in each step. We show that the resulting block update problems can be solved in closed form even when the structural equation model comprises feedback cycles. Furthermore, we give a characterization of the models for which the block-coordinate descent algorithm is well defined, meaning that for generic data and starting values all block optimization problems admit a unique solution. For the characterization, we represent each model by its mixed graph (also known as path diagram), which leads to criteria that can be checked in time that is polynomial in the number of considered variables.

1. Introduction. Structural equation models (SEMs) provide a general framework for modeling stochastic dependence that arises through cause-effect relationships between random variables. The models form a cornerstone of multivariate statistics with applications ranging from biology to the social sciences [Bollen (1989), Hoyle (2012), Kline (2015)]. Through their representation by path diagrams, which originate in the work of Wright (1921, 1934), the models encompass directed graphical models [Lauritzen (1996)]. While SEMs can naturally be interpreted as models of causality that predict effects of experimental interventions [Pearl (2009), Spirtes, Glymour and Scheines (2000)], the focus of this paper is on observational scenarios. In other words, we consider statistical inference based on independent and identical samples from a distribution in an SEM. Concretely, we will treat linear SEMs in which the effects of any latent variables are marginalized out and represented through correlation among the error terms in the structural equations; see, for example, Pearl [(2009), Section 3.7], Spirtes, Glymour and Scheines [(2000), Chapter 6] or Wermuth (2011). This setting arises, in particular, in problems of network recovery through model selection as treated, for exam-

Received October 2016; revised May 2017.
MSC2010 subject classifications. 62H12, 62F10.
Key words and phrases. Cyclic graph, feedback, linear structural equation model, graphical model, maximum likelihood estimation.

ple, by Colombo et al. (2012), Silva (2013), Nowzohour, Maathuis and Bühlmann (2015) or Triantafillou and Tsamardinos (2016). For further details and references, see Section 5.2 in Drton and Maathuis (2017).

The specific problem we address is the computation of maximum likelihood estimates (MLEs) in linear SEMs with Gaussian errors in the structural equations. The R packages "sem" [Fox (2006)] and "lavaan" [Rosseel (2012)] as well as commercial software [Narayanan (2012)] solve this problem by applying general quasi-Newton methods for nonlinear optimization. However, these methods are often subject to convergence problems and may require careful choice of starting values [Steiger (2001)]. This is particularly exacerbated when computing MLEs in poorly fitting models as part of model selection [Drton, Eichler and Richardson (2009)]. As a software manual puts it: "It can be devilishly difficult for software to obtain results for SEMs" [StataCorp (2013), page 112].

As an alternative, we propose a block-coordinate descent (BCD) method that cycles through the considered variables, updating the parameters related to a given variable in each step. Each update is performed through partial maximization of the likelihood function. This method generalizes the iterative conditional fitting algorithm of Chaudhuri, Drton and Richardson (2007) as well as the algorithm of Drton, Eichler and Richardson (2009). In contrast to this earlier work, our extension is applicable to models that comprise feedback cycles. Models with feedback cycles have been treated by Spirtes (1995), Richardson (1996, 1997), and more recently by Lacerda et al. (2008), Mooij and Heskes (2013) and Park and Raskutti (2016). An example of a recent application can be found in the work of Grace et al. (2016).

The presence of feedback loops complicates likelihood inference as even in settings without latent variables MLEs are generally high-degree algebraic functions of the data. For example, the MLE in the model given by the graph in Figure 1 is an algebraic function of degree 7; see Chapter 2.1 in Drton, Sturmfels and Sullivant (2009) for how to compute this degree. Somewhat surprisingly, however, the update steps in our BCD algorithm admit a closed form even in the presence of feedback loops, and the computational effort is on the same order as in the case without feedback loops. In numerical experiments, the BCD algorithm is seen to avoid convergence problems.

As a second main contribution, we show that the algorithm applies to interesting models with "bows." In terms of the mixed graph/path diagram, a bow is a subgraph on two nodes i and j with two edges $i \to j$ and $i \leftrightarrow j$. Such a subgraph indicates that there is both a direct effect of the ith variable on the jth variable as well as a latent confounder with effects on the two variables. Bows can lead to

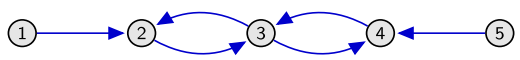

FIG. 1. *Graph of a cyclic linear SEM with maximum likelihood degree* 7.

collinearity issues in the BCD algorithm, and we are able to give a characterization of the models for which the algorithm is well defined, meaning that for generic data and starting values all block optimization problems admit a unique and feasible solution. For the characterization, we represent each model by its mixed graph/path diagram, which leads to criteria that can be checked in time that is polynomial in the number of considered variables.

The paper is organized as follows. In Section 2, we review necessary background on SEMs. The new BCD algorithm is derived in Section 3. Its properties are discussed in Section 4. Numerical examples are presented in Section 5. We conclude with a discussion of the considered problem in Section 6.

2. Linear structural equation models.

2.1. *Basics.* A structural equation model (SEM) captures dependence among a set of variables $\{Y_i : i \in V\}$. Each model is built from a system of equations, with one equation for each considered variable. Each such *structural equation* specifies how a variable Y_i arises as a function of the other variables and a stochastic error term ε_i. In the linear case considered here, we have

$$Y_i = \sum_{j \in V \setminus \{i\}} \beta_{ij} Y_j + \varepsilon_i, \qquad i \in V. \tag{2.1}$$

Collecting the Y_i and ε_i terms into the vectors Y and ε, respectively, (2.1) can be rewritten as

$$Y = BY + \varepsilon, \tag{2.2}$$

where $B = (\beta_{ij})$ is a matrix of coefficients that are sometimes termed *structural parameters* [Bollen (1989)]. Specific models of interest are obtained by assuming that for some index pairs (i, j), variable Y_j has no direct effect on Y_i, which in the linear framework is encoded by the restriction that $\beta_{ij} = 0$.

Techniques for statistical inference are often based on the assumption that ε follows a multivariate normal distribution with possible dependence among its coordinates. So,

$$\varepsilon \sim \mathcal{N}(0, \Omega), \tag{2.3}$$

where $\Omega = (\omega_{ij})$ is a symmetric, positive definite matrix of parameters. An entry ω_{ij} may capture effects of potential latent variables that are common causes of Y_i and Y_j. When no latent common cause of Y_i and Y_j is believed to exist, constrain $\omega_{ij} = \omega_{ji} = 0$ [see, e.g., Pearl (2009), Spirtes, Glymour and Scheines (2000)]. As a result of (2.2) and (2.3), the observed random variables, Y, have a centered normal distribution with covariance matrix

$$\Sigma = (I - B)^{-1} \Omega (I - B)^{-T}. \tag{2.4}$$

Here, I is the $V \times V$ identity matrix. Note that the assumption of centered variables can be made without loss of generality [Anderson (2003), Chapter 7].

It is often convenient to represent an SEM by a mixed graph or path diagram [Wright (1921, 1934)]. The graph has vertex set V and is mixed in the sense of having both a set of *directed edges* $E_{\rightarrow}$ and a set of *bi-directed edges* $E_{\leftrightarrow}$. The directed edges in $E_{\rightarrow}$ are ordered pairs in $V \times V$, whereas the edges in $E_{\leftrightarrow}$ have no orientation and are unordered pairs $\{i, j\}$ with $i, j \in V$. We will often write $i \rightarrow j$ in place of (i, j) for a potential edge in $E_{\rightarrow}$ and $i \leftrightarrow j$ for a potential edge $\{i, j\}$ in $E_{\leftrightarrow}$. In this setup, each variable Y_i is then represented by a node, corresponding to its index $i \in V$. An edge $j \rightarrow i$ is not in $E_{\rightarrow}$ if and only if the model imposes the constraint that $\beta_{ij} = 0$. Note that in our context there are no self-loops $i \rightarrow i$. Similarly, the edge $i \leftrightarrow j$ is absent from $E_{\leftrightarrow}$ if and only if the model imposes the constraint that $\omega_{ij} = \omega_{ji} = 0$. Finally, for each node $j \in V$, we define two sets $\mathrm{pa}(j)$ and $\mathrm{sib}(j)$ that we refer to as the *parents* and *siblings* of j, respectively. The set $\mathrm{pa}(j)$ comprises all nodes $i \in V$ such that $i \rightarrow j \in E_{\rightarrow}$, and $\mathrm{sib}(j)$ is the set of all nodes $i \in V$ such that $i \leftrightarrow j \in E_{\leftrightarrow}$.

Let $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be a mixed graph, and define $\mathbf{B}(G)$ to be the set of real $V \times V$ matrices $B = (\beta_{ij})$ such that $I - B$ is invertible and

$$\beta_{ij} = 0 \qquad \text{whenever } j \rightarrow i \notin E_{\rightarrow}. \tag{2.5}$$

Similarly, define $\mathbf{\Omega}(G)$ to be the set of all positive definite symmetric $V \times V$ matrices $\Omega = (\omega_{ij})$ that satisfy

$$\omega_{ij} = 0 \qquad \text{whenever } j \leftrightarrow i \notin E_{\leftrightarrow}. \tag{2.6}$$

The *linear SEM* $\mathbf{N}(G)$ associated with graph G is then the family of multivariate normal distributions $\mathcal{N}(0, \Sigma)$ with covariance matrix Σ as in (2.4) for $B \in \mathbf{B}(G)$ and $\Omega \in \mathbf{\Omega}(G)$.

A mixed graph G and the associated model $\mathbf{N}(G)$ are *cyclic* if G contains a directed cycle, that is, a subgraph of the form

$$i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_k \rightarrow i_1$$

for distinct nodes $i_1, \ldots i_k \in V$, $k \geq 2$. If there is no such cycle, the graph and corresponding model are said to be *acyclic*. Acyclicity brings about great simplifications as we have $\det(I - B) = 1$ for every $B \in \mathbf{B}(G)$ if and only if G is acyclic. To see this note that when G is acyclic, there exists a topological ordering of V, that is, a relabeling of V such that $i \rightarrow j \in E_{\rightarrow}$ only if $i < j$. Under such an ordering every matrix in $\mathbf{B}(G)$ is strictly lower triangular. If G is an acyclic digraph, such that $E_{\leftrightarrow} = \varnothing$ then the MLE in $\mathbf{N}(G)$ is obtained by solving a linear regression problem for each variable Y_i, $i \in V$. For an acyclic graph with $E_{\leftrightarrow} \neq \varnothing$, this is generally no longer the case but the MLE can be found by iterative least squares computations [Drton, Eichler and Richardson (2009)].

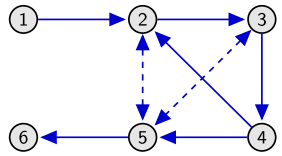

FIG. 2. *Cyclic mixed graph that is almost everywhere identifiable.*

2.2. *Cyclic models.* A challenge in the computation of MLEs in models with cyclic path diagrams is the fact that $\det(I - B)$ is not constant one. For example, $\det(I - B) = 1 - \beta_{32}\beta_{43}\beta_{24}$ for matrices $B \in \mathbf{B}(G)$ when G is the mixed graph in Figure 2. We observe a correspondence between the term $\beta_{32}\beta_{43}\beta_{24}$ and the directed cycle $2 \to 3 \to 4 \to 2$ in the graph. We now review this connection in the setting of a general mixed graph G.

Let $\mathbf{S}_V$ be the symmetric group of all permutations of the vertex set V. Every permutation $\sigma \in \mathbf{S}_V$ has a unique decomposition into disjoint permutation cycles. Let $\mathcal{C}(\sigma)$ be the set of permutation cycles of σ, and let $\mathcal{C}_2(\sigma)$ be the subset containing cycles of length 2 or more. Write $n(\sigma)$ for the cardinality of $\mathcal{C}_2(\sigma)$, and $V(\sigma)$ for the set of nodes that are contained in a cycle in $\mathcal{C}_2(\sigma)$. Moreover, define

$$\mathbf{S}_V(G) = \big\{\sigma \in \mathbf{S}_V : i = \sigma(i) \text{ or } i \to \sigma(i) \in E_{\to} \text{ for all } i \in V\big\}. \tag{2.7}$$

LEMMA 1. *Let $B = (\beta_{ij}) \in \mathbf{B}(G)$ for a mixed graph $G = (V, E_{\to}, E_{\leftrightarrow})$. Then*

$$\det(I - B) = \sum_{\sigma \in \mathbf{S}_V(G)} (-1)^{n(\sigma)} \prod_{i \in V(\sigma)} \beta_{\sigma(i),i}.$$

The lemma follows from a Leibniz expansion of the determinant. It could be derived from Theorem 1 in Harary (1962) by treating the diagonal of $I - B$ as self-loops with weight 1 and taking into account that B is negated. We include its proof in the Supplementary Material [Drton, Fox and Wang (2018)].

When deriving the block-coordinate descent algorithm proposed in Section 3, we treat $\det(I - B)$ as a function of only the entries in a given row. By multilinearity of the determinant, this function is linear and its coefficients are obtained in a Laplace expansion. Throughout the paper, we let $-i := V \setminus \{i\}$ and denote the $U \times W$ submatrix of a matrix A by $A_{U,W}$.

LEMMA 2. *Let $B = (\beta_{ij}) \in \mathbf{B}(G)$ for a mixed graph $G = (V, E_{\to}, E_{\leftrightarrow})$. Fix an arbitrary node $i \in V$. Then $\det(I - B)$ is linear in the entries of $B_{i,\mathrm{pa}(i)} = (\beta_{ij} : j \in \mathrm{pa}(i))$ with*

$$\det(I - B) = c_{i,0} + B_{i,\mathrm{pa}(i)} c_{i,\mathrm{pa}(i)},$$

where $c_{i,0} \in \mathbb{R}$ and the entries of $c_{i,\mathrm{pa}(i)} \in \mathbb{R}^{\mathrm{pa}(i)}$ are subdeterminants, namely,

$$c_{i,0} = \det\big((I - B)_{-i,-i}\big),$$

$$c_{i,p} = (-1)^{i+p-1} \det\big((I - B)_{-i,-p}\big), \qquad p \in \mathrm{pa}(i);$$

to define $(-1)^{i+p-1}$ *enumerate* V *in accordance with the layout of the matrix* $I - B$.

EXAMPLE 1. The mixed graph G from Figure 2 encodes the equation system

$$\begin{aligned} Y_1 &= \varepsilon_1, \qquad & Y_2 &= \beta_{21}Y_1 + \beta_{24}Y_4 + \varepsilon_2, \\ Y_3 &= \beta_{32}Y_2 + \varepsilon_3, \qquad & Y_4 &= \beta_{43}Y_3 + \varepsilon_4, \\ Y_5 &= \beta_{54}Y_4 + \varepsilon_5, \qquad & Y_6 &= \beta_{65}Y_5 + \varepsilon_6, \end{aligned}$$

where ε_1, ε_2, ε_3, ε_4 and ε_6 are all pairwise uncorrelated, and ε_5 is uncorrelated with ε_1, ε_4 and ε_6. The system contains the directed cycle $2 \to 3 \to 4 \to 2$. Consequently,

$$\det(I - B) = 1 - \beta_{32}\beta_{43}\beta_{24}.$$

Hence, the coefficients must satisfy $\beta_{32}\beta_{43}\beta_{24} \neq 1$ for the equation system to yield a positive definite covariance matrix. When fixing node $i \in V$ and writing $\det(I - B)$ as a linear function of $(\beta_{ij})_{j \in \mathrm{pa}(i)}$ as in Lemma 2, we have

$$\begin{aligned} &c_{1,0} = 1 - \beta_{32}\beta_{43}\beta_{24}, && \mathrm{pa}(1) = \varnothing; \\ &c_{2,0} = 1, && \mathrm{pa}(2) = \{1, 4\}, && c_{2,\mathrm{pa}(2)} = (0, -\beta_{32}\beta_{43})^T; \\ &c_{3,0} = 1, && \mathrm{pa}(3) = \{2\}, && c_{3,\mathrm{pa}(3)} = -\beta_{43}\beta_{24}; \\ &c_{4,0} = 1, && \mathrm{pa}(4) = \{3\}, && c_{4,\mathrm{pa}(4)} = -\beta_{32}\beta_{24}; \\ &c_{5,0} = 1 - \beta_{32}\beta_{43}\beta_{24}, && \mathrm{pa}(5) = \{4\}, && c_{5,\mathrm{pa}(5)} = 0; \\ &c_{6,0} = 1 - \beta_{32}\beta_{43}\beta_{24}, && \mathrm{pa}(6) = \{5\}, && c_{6,\mathrm{pa}(6)} = 0. \end{aligned}$$

2.3. *Likelihood inference.* Suppose we are given a sample of N observations in $\mathbb{R}^V$. Let Y be the $V \times N$ matrix with these observations as columns, and let $S = \frac{1}{N}YY^T$ be the associated $V \times V$ sample covariance matrix (for known zero mean). Fix a possibly cyclic mixed graph G. Ignoring an additive constant and dividing out a factor of $N/2$, model $\mathbf{N}(G)$ has log-likelihood function

$$\begin{aligned} \ell_{G,Y}(\Omega, B) = &- \log\det(\Omega) + \log\det(I - B)^2 \\ &- \mathrm{tr}\big\{(I - B)^T\Omega^{-1}(I - B)S\big\}. \end{aligned} \tag{2.8}$$

Throughout the paper, we assume that Y has full rank $|V|$. This holds with probability one if the sample is from a continuous distribution and $N \geq |V|$. Full rank of Y implies that S is positive definite, and the log-likelihood function $\ell_{G,Y}$ is then bounded for any graph G. However, if G is sparse with a bi-directed part $(V, E_{\leftrightarrow})$ that is not connected, then $\ell_{G,Y}$ may also be bounded if S is not positive definite [Fox (2014)].

Our problem of interest is to compute (local) maxima of the log-likelihood function. These solve the likelihood equations, which are obtained by equating to zero the gradient of $\ell_{G,Y}(B, \Omega)$. To be precise, the partial derivatives are taken with respect to the free entries in B and Ω, which we denote by β and ω, respectively. So, β has $|E_{\to}|$ entries, and ω has $|V| + |E_{\leftrightarrow}|$ entries. Let $\operatorname{vec}(A)$ denote the vectorization (stacking of the columns) of a matrix A. Then there are 0/1-valued matrices P and Q such that $\operatorname{vec}(B) = P\beta$, and $\operatorname{vec}(\Omega) = Q\omega$.

PROPOSITION 1. *The likelihood equations of the model* $\mathbf{N}(G)$ *can be written as*

$$P^T \operatorname{vec}\big[\Omega^{-1}(I - B)S - (I - B)^T\big] = 0, \tag{2.9}$$

$$Q^T \operatorname{vec}\big(\Omega^{-1} - \Omega^{-1}(I - B)S(I - B)^T\Omega^{-1}\big) = 0. \tag{2.10}$$

A derivation of this result is provided in the Supplementary Material [Drton, Fox and Wang (2018)]. In general, the likelihood equations are difficult to solve analytically; recall the example from Figure 1. Instead, it is common practice to use iterative maximization techniques.

3. Block-coordinate descent for cyclic mixed graphs.

3.1. *Algorithm overview.* We now introduce our block-coordinate descent (BCD) procedure for computing the MLE in a possibly cyclic mixed graph model $\mathbf{N}(G)$. The method requires initializing with a choice of $B \in \mathbf{B}(G)$ and $\Omega \in \mathbf{\Omega}(G)$. The algorithm then proceeds by repeatedly iterating through all nodes in V and performing update steps. In the update for node i, we maximize the log-likelihood function with respect to all parameters corresponding to edges with a head at i (i.e., $B_{i,\mathrm{pa}(i)}$ and $\Omega_{i,\mathrm{sib}(i)\cup\{i\}}$) while holding all other parameters fixed. The parameters that are updated determine the ith row in B and the ith row and column in the symmetric matrix Ω. The algorithm stops when a convergence criterion is satisfied.

In the derivation of the block update, we write Y_C for the $C \times N$ submatrix of Y, for subset $C \subset V$. In particular, $Y_{-i} = Y_{V\setminus\{i\}}$, and Y_i is the ith row of Y. Finally, we will invoke assumptions to ensure that the optimization problem yielding the block update admits a unique solution. The graphs G for which these assumptions hold will be characterized in Section 4.

3.2. *Block update problem.* In the ith block update problem, we seek to maximize the log-likelihood function $\ell_{G,Y}$ while holding the submatrices $\Omega_{-i,-i}$ and B_{-i} fixed. Let

$$\omega_{ii.-i} = \omega_{ii} - \Omega_{i,-i}\Omega_{-i,-i}^{-1}\Omega_{-i,i}$$

be the conditional variance of the error term ε_i given ε_{-i}; here, $\Omega^{-1}_{-i,-i} = (\Omega_{-i,-i})^{-1}$. In analogy to Theorem 12 in Drton, Eichler and Richardson (2009), the log-likelihood function can be decomposed as

$$\begin{aligned}
&\ell_{G,Y}(\Omega, B)\\
&\quad = -\log \omega_{ii.-i}\\
&\qquad - \frac{1}{N\omega_{ii.-i}} \|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} \big(\Omega^{-1}_{-i,-i}\varepsilon_{-i}\big)_{\mathrm{sib}(i)}\|^2\\
&\qquad - \log\det(\Omega_{-i,-i}) - \frac{1}{N}\operatorname{tr}\big(\Omega^{-1}_{-i,-i}\varepsilon_{-i}\varepsilon^T_{-i}\big) + \log\det(I-B)^2.
\end{aligned} \tag{3.1}$$

This follows by factoring the joint distribution of ε into the marginal distribution of ε_{-i} and the conditional distribution of ε_i given ε_{-i}. The key difference between (3.1) and the corresponding log-likelihood decomposition in Drton, Eichler and Richardson (2009) is the presence of the term $\log\det(I-B)^2$, which is nonzero for cyclic graphs.

With $\Omega_{-i,-i}$ and B_{-i} fixed, we can first compute the error terms

$$\varepsilon_{-i} = (I-B)_{-i} Y \tag{3.2}$$

and subsequently the *pseudo-variables*

$$Z_{-i} = \Omega^{-1}_{-i,-i}\varepsilon_{-i}. \tag{3.3}$$

From (3.1), it is clear that, for fixed $\Omega_{-i,-i}$ and B_{-i}, the maximization of $\ell_{G,Y}$ reduces to the maximization of the function

$$\begin{aligned}
&\ell_{G,Y,i}(\Omega_{i,\mathrm{sib}(i)}, \omega_{ii.-i}, B_{i,\mathrm{pa}(i)})\\
&\quad = -\log\omega_{ii.-i} + \log\big[(c_{i,0} + B_{i,\mathrm{pa}(i)} c_{i,\mathrm{pa}(i)})^2\big]\\
&\qquad - \frac{1}{N\omega_{ii.-i}} \|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)}\|^2.
\end{aligned} \tag{3.4}$$

Here, we applied Lemma 2, and let $B_{i,\mathrm{pa}(i)} = (\beta_{ij} : j \in \mathrm{pa}(i))$ and $\Omega_{i,\mathrm{sib}(i)} = (\omega_{ik} : k \in \mathrm{sib}(i))$. The domain of definition of $\ell_{G,Y,i}$ is $\mathbb{R}^{\mathrm{sib}(i)} \times (0,\infty) \times \mathbb{R}^{\mathrm{pa}(i)}_{\mathrm{inv}}$, where

$$\mathbb{R}^{\mathrm{pa}(i)}_{\mathrm{inv}} = \mathbb{R}^{\mathrm{pa}(i)} \setminus \{B_{i,\mathrm{pa}(i)} : c_{i,0} + B_{i,\mathrm{pa}(i)} c_{i,\mathrm{pa}(i)} = 0\}$$

excludes choices of $B_{i,\mathrm{pa}(i)}$ for which $I-B$ is noninvertible. For any fixed choice of $B_{i,\mathrm{pa}(i)}$ and $\Omega_{i,\mathrm{sib}(i)}$, if $Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)} \neq 0$ then

$$\omega^{\star}_{ii.-i} = \frac{1}{N}\|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)}\|^2 \tag{3.5}$$

uniquely maximizes $\ell_{G,Y,i}$ with respect to $\omega_{ii.-i}$. This fact could be used to form a profile log-likelihood function. Before proceeding, however, we shall address the concern that for a mixed graph G that contains cycles, it may occur that

$Y_i \in \operatorname{span}(Y_{\operatorname{pa}(i)}, Z_{\operatorname{sib}(i)})$ even if the rows of Y are linearly independent. A simple example would be the graph with nodes 1 and 2 and three edges $1 \to 2$, $1 \leftarrow 2$ and $1 \leftrightarrow 2$; see Example 4 below.

LEMMA 3. *Let the data matrix $Y \in \mathbb{R}^{V \times N}$ have linearly independent rows. Then $Y_i - B_{i,\operatorname{pa}(i)} Y_{\operatorname{pa}(i)} - \Omega_{i,\operatorname{sib}(i)} Z_{\operatorname{sib}(i)} \neq 0$ for all $B \in \mathbf{B}(G)$, $\Omega \in \mathbf{\Omega}(G)$ and $i \in V$.*

PROOF. From (3.3),

$$Y_i - B_{i,\operatorname{pa}(i)} Y_{\operatorname{pa}(i)} - \Omega_{i,\operatorname{sib}(i)} Z_{\operatorname{sib}(i)} = \varepsilon_i - \omega_{i,-i} \Omega_{-i,-i}^{-1} \varepsilon_{-i} = 0$$

only if $\varepsilon = (I - B)Y \in \mathbb{R}^{V \times N}$ has linearly dependent rows. However, this cannot occur when Y has linearly independent rows as matrices $B \in \mathbf{B}(G)$ have $I - B$ invertible. □

According to Lemma 3, we may indeed substitute $\omega^{\star}_{ii.-i}$ from (3.5) into $\ell_{G,Y,i}$ and maximize the resulting profile log-likelihood function:

$$\begin{aligned}(\Omega_{i,\operatorname{sib}(i)}, B_{i,\operatorname{pa}(i)}) \mapsto{}& \log(N) - 1 \\ &- \log\biggl(\frac{\|Y_i - B_{i,\operatorname{pa}(i)} Y_{\operatorname{pa}(i)} - \Omega_{i,\operatorname{sib}(i)} Z_{\operatorname{sib}(i)}\|^2}{(c_{i,0} + B_{i,\operatorname{pa}(i)} c_{i,\operatorname{pa}(i)})^2}\biggr).\end{aligned} \tag{3.6}$$

By monotonicity of the logarithm, maximizing (3.6) with respect to the structural parameters $(\Omega_{i,\operatorname{sib}(i)}, B_{i,\operatorname{pa}(i)}) \in \mathbb{R}^{\operatorname{sib}(i)} \times \mathbb{R}^{\operatorname{pa}(i)}_{\operatorname{inv}}$ is equivalent to minimizing

$$g_i(\Omega_{i,\operatorname{sib}(i)}, B_{i,\operatorname{pa}(i)}) = \frac{\|Y_i - B_{i,\operatorname{pa}(i)} Y_{\operatorname{pa}(i)} - \Omega_{i,\operatorname{sib}(i)} Z_{\operatorname{sib}(i)}\|^2}{(c_{i,0} + B_{i,\operatorname{pa}(i)} c_{i,\operatorname{pa}(i)})^2}. \tag{3.7}$$

If $c_{i,\operatorname{pa}(i)} = 0$, which occurs when i does not lie on any directed cycle, then the denominator in (3.7) is constant and the problem amounts to finding least squares estimates for $\Omega_{i,\operatorname{sib}(i)}$ and $B_{i,\operatorname{pa}(i)}$. In other words, we solve a linear regression problem with response Y_i and covariates Z_k, $k \in \operatorname{sib}(i)$ and Y_j, $j \in \operatorname{pa}(i)$. This is the setting of Drton, Eichler and Richardson (2009).

In the more difficult case where $c_{i,\operatorname{pa}(i)} \neq 0$, minimizing the function g_i from (3.7) amounts to minimizing a ratio of two univariate quadratic functions. The numerator is a least squares objective for a linear regression problem with design matrix $(Z^T_{\operatorname{sib}(i)}, Y^T_{\operatorname{pa}(i)}) \in \mathbb{R}^{N \times (|\operatorname{sib}(i)| + |\operatorname{pa}(i)|)}$. The denominator is the square of an affine function whose slope vector satisfies the following property proven in the Supplementary Material [Drton, Fox and Wang (2018)].

LEMMA 4. *The vector $\begin{pmatrix} 0 \\ c_{i,\operatorname{pa}(i)} \end{pmatrix}$ is orthogonal to the kernel of $\begin{pmatrix} Z_{\operatorname{sib}(i)} \\ Y_{\operatorname{pa}(i)} \end{pmatrix}^T$.*

3.3. *Minimizing a ratio of quadratic functions.* When $c_{i,\mathrm{pa}(i)} \neq 0$, the minimization of g_i from (3.7) is an instance of the general problem

$$\min_{\alpha \in \mathbb{R}^m} \frac{\|y - X\alpha\|^2}{(c_0 + c^T\alpha)^2} \tag{3.8}$$

that is specified by a vector $y \in \mathbb{R}^N$ with $N \geq m$, a matrix $X \in \mathbb{R}^{N \times m}$, a nonzero vector $c \in \mathbb{R}^m \setminus \{0\}$ and a scalar $c_0 \in \mathbb{R}$. For a correspondence to (3.7), take as argument the vector $\alpha = (\Omega_{i,\mathrm{sib}(i)}, B_{i,\mathrm{pa}(i)})^T$, which is of length $m = |\,\mathrm{sib}(i)| + |\,\mathrm{pa}(i)|$, and set

$$y = Y_i^T, \qquad X = \begin{pmatrix} Z_{\mathrm{sib}(i)} \\ Y_{\mathrm{pa}(i)} \end{pmatrix}^T, \qquad c = \begin{pmatrix} 0 \\ c_{i,\mathrm{pa}(i)} \end{pmatrix}, \qquad c_0 = c_{i,0}. \tag{3.9}$$

We now show that (3.8) admits a closed-form solution. In doing so, we focus attention on problems in which the matrix X has full column rank. Unless stated otherwise, we do not require that c be orthogonal to the kernel of X. Rank deficient cases are discussed in Remark 2 at the end of this section.

THEOREM 1. *Suppose the matrix X has full rank $m \leq N$. Let $\hat{\alpha} = (X^T X)^{-1} \times X^T y$ be the minimizer $\alpha \mapsto \|y - X\alpha\|^2$, and let $y_0^2 = \|y - X\hat{\alpha}\|^2$:*

(i) *If $c_0 + c^T\hat{\alpha} \neq 0$, then (3.8) is uniquely solved by*

$$\alpha^\star = \hat{\alpha} + \frac{y_0^2}{c_0 + c^T\hat{\alpha}}(X^T X)^{-1}c.$$

(ii) *If $c_0 + c^T\hat{\alpha} = 0$ and $y_0^2 = 0$, then (3.8) admits a solution, but not uniquely so. The solution set is $\{\hat{\alpha} + \lambda(X^T X)^{-1}c : \lambda \in \mathbb{R} \setminus \{0\}\}$.*

(iii) *If $c_0 + c^T\hat{\alpha} = 0$ and $y_0^2 > 0$, the minimum in (3.8) is not achieved.*

REMARK 1. The computational complexity of solving (3.8) is on the same order as that for the least squares problem with objective $\|y - X\alpha\|^2$.

PROOF OF THEOREM 1. We give a numerically stable algorithm for solving (3.8), and then translate the solution into a rational function of the input (y, X, c_0, c).

(a) *Algorithm.* Find an orthogonal $m \times m$ matrix Q_1 such that $Q_1 c = (0, \ldots, 0, \|c\|)^T$; note that in our context the support of c is confined to the coordinates indexed by $\mathrm{pa}(i)$. Reparametrizing to $\alpha' = Q_1\alpha$, (3.8) becomes

$$\min_{\alpha' \in \mathbb{R}^m} \frac{\|y - XQ_1^T\alpha'\|^2}{(c_0 + \|c\|\alpha'_m)^2} \tag{3.10}$$

with α'_m being the last coordinate of $\alpha' = (\alpha'_1, \ldots, \alpha'_m)$. Next, compute a QR decomposition $XQ_1^T = Q_2^T R$, where Q_2 is an orthogonal $N \times N$ matrix, and R is

an upper triangular $N \times m$ matrix. Observe that $R = \binom{R_1}{0}$ with $R_1 \in \mathbb{R}^{m\times m}$ upper triangular. Since orthogonal transformations leave Euclidean norms invariant,

$$\frac{\|y - XQ_1^T\alpha'\|^2}{(c_0 + \|c\|\alpha'_m)^2} = \frac{\|Q_2 y - R\alpha'\|^2}{(c_0 + \|c\|\alpha'_m)^2} = \frac{\sum_{j=1}^m [(Q_2y)_j - (R_1\alpha')_j]^2 + y_0^2}{(c_0 + \|c\|\alpha'_m)^2}, \tag{3.11}$$

where $y_0^2 = \sum_{j=m+1}^N (Q_2y)_j^2$ is the squared length of the projection of y on the orthogonal complement of the span of X. Finally, we reparametrize to $\alpha'' = R_1\alpha'$ and obtain the problem

$$\min_{\alpha''\in\mathbb{R}^m} \frac{\sum_{j=1}^m [(Q_2y)_j - \alpha''_j]^2 + y_0^2}{(c_0 + \|c\|r^{-1}\alpha''_m)^2} \tag{3.12}$$

with $r = R_{mm}$ being the (m, m) entry in R (and R_1). We have $r \neq 0$ as X and thus XQ_1^T and also R have full column rank. This also entails that R_1 is invertible.

For α'' to be a solution of (3.12), it clearly must hold that

$$\alpha''_j = (Q_2y)_j \qquad \text{for } j = 1, \dots, m-1, \tag{3.13}$$

and (3.12) is solved by finding the coordinate α''_m by minimizing the univariate function

$$g(\alpha''_m) = \frac{((Q_2y)_m - \alpha''_m)^2 + y_0^2}{(c_0 + \|c\|r^{-1}\alpha''_m)^2}, \qquad \alpha''_m \in \mathbb{R}. \tag{3.14}$$

By Lemma 5 below and assuming that $c_0 + \|c\|r^{-1}(Q_2y)_m \neq 0$, the univariate function g from (3.14) attains its minimum at

$$\alpha''_m = (Q_2y)_m + \frac{\|c\|y_0^2}{rc_0 + \|c\|(Q_2y)_m}. \tag{3.15}$$

If $c_0 + \|c\|r^{-1}(Q_2y)_m = 0$ and $y_0^2 = 0$, then g is constant and any feasible $\alpha''_m \neq (Q_2y)_m$ is optimal. If $c_0 + \|c\|r^{-1}(Q_2y)_m = 0$ and $y_0^2 > 0$, then g does not achieve its minimum.

In order to solve the problem posed at the beginning of this subsection, that is, the problem from (3.8), we convert the optimum α'' from (3.13) and (3.15) to

$$\alpha = Q_1^T R_1^{-1}\alpha''. \tag{3.16}$$

(b) *Rational formulas.* Inspecting (3.11), we observe that $R_1^{-1}(Q_2y)_{\{1,\dots,m\}}$ is the coefficient vector that solves the least squares problem in which y is regressed on XQ_1^T. Therefore,

$$Q_1^T R_1^{-1}(Q_2y)_{\{1,\dots,m\}} = (X^TX)^{-1}X^Ty =: \hat{\alpha} \tag{3.17}$$

is the least squares coefficient vector for the regression of y on X. Because R_1 is rectangular, it follows that $r^{-1}(Q_2y)_m$ is the mth entry of the vector

$R_1^{-1}(Q_2 y)_{\{1,\dots,m\}}$. With $Q_1 c = (0,\dots,0,\|c\|)^T$, we deduce that

$$\begin{aligned}\|c\| r^{-1}(Q_2 y)_m &= \langle Q_1 c, R_1^{-1}(Q_2 y)_{\{1,\dots,m\}}\rangle = \langle c, Q_1^T R^{-1}(Q_2 y)_{\{1,\dots,m\}}\rangle \\ &= \langle c, \hat{\alpha}\rangle.\end{aligned} \tag{3.18}$$

Let $e_m = (0,\dots,0,1)^T$ be the mth canonical basis vector. Using that R_1^{-T} has its last column equal to $r^{-1}e_m$, we find that

$$\begin{aligned}Q_1^T R_1^{-1} e_m \|c\| r^{-1} &= Q_1^T R_1^{-1} R_1^{-T} Q_1 c = (Q_1^T R^T R Q_1)^{-1} c \\ &= (Q_1^T R^T Q_2 Q_2^T R Q_1)^{-1} c = (X^T X)^{-1} c.\end{aligned} \tag{3.19}$$

In case (i), we obtain from (3.13), (3.15) and (3.16) that the unique minimum is

$$Q_1^T R_1^{-1}(Q_2 y)_{\{1,\dots,m\}} + \frac{\|c\| r^{-1} y_0^2}{c_0 + \|c\| r^{-1}(Q_2 y)_m} Q_1^T R_1^{-1} e_m.$$

Applying (3.17)–(3.19), we readily find the rational formula asserted in the theorem. Cases (ii) and (iii) are similar. □

The proof used the following lemma about ratios of univariate quadratics. The lemma is derived in the Supplementary Material [Drton, Fox and Wang (2018)].

LEMMA 5. *For constants $a, b, c_0, c_1 \in \mathbb{R}$ with $c_1 \neq 0$, define the function*

$$f(x) = \frac{(a-x)^2 + b^2}{(c_0 + c_1 x)^2}, \qquad x \in \mathbb{R} \setminus \{-c_0/c_1\}.$$

(i) *If $c_0 + a c_1 \neq 0$, then f is uniquely minimized by*

$$x = \frac{a c_0 + a^2 c_1 + b^2 c_1}{c_0 + a c_1} = a + \frac{b^2 c_1}{c_0 + a c_1}.$$

(ii) *If $c_0 + a c_1 = 0$ and $b = 0$, then f is constant and equal to $1/c_1^2$.*

(iii) *If $c_0 + a c_1 = 0$ and $b^2 > 0$, then f does not achieve its minimum, and* $\inf f = \lim_{x \to \pm\infty} f(x) = 1/c_1^2$.

REMARK 2. When c is orthogonal to the kernel of X, then $c = X^T \tilde{c}$ for a vector $\tilde{c} \in \mathbb{R}^N$. The problem (3.8) is then equivalent to

$$\min_{\tilde{\alpha} \in \operatorname{span}(X)} \frac{\|y - \tilde{\alpha}\|^2}{(c_0 + \tilde{c}^T \tilde{\alpha})^2}. \tag{3.20}$$

Let $\mathcal{L}(X)$ be the column span of X, and let $\pi_{\mathcal{L}(X)}$ be the orthogonal projection onto $\mathcal{L}(X)$. Then (3.20) admits a unique solution if and only if $c_0 + \tilde{c}^T \pi_{\mathcal{L}(X)}(y) \neq 0$. The unique solution is

$$\tilde{\alpha}^\star = \pi_{\mathcal{L}(X)}(y) + \frac{\|y - \pi_{\mathcal{L}(X)}(y)\|^2}{c_0 + \tilde{c}^T \pi_{\mathcal{L}(X)}(y)} \pi_{\mathcal{L}(X)}(\tilde{c}),$$

which is meaningful also when X does not have full rank. If desired, a coefficient vector $\alpha^\star \in \mathbb{R}^m$ satisfying $X\alpha^\star = \tilde{\alpha}^\star$ can be chosen.

3.4. *The BCD algorithm.* By Theorem 1, or rather the algorithm outlined in its proof, we are able to efficiently minimize the function g_i from (3.7). In other words, we can efficiently update the ith row in B and the ith row and column in Ω by a partial maximization of the log-likelihood function $\ell_{G,Y}$. We summarize our block-coordinate descent scheme for maximization of the log-likelihood function $\ell_{G,Y}$ in Algorithm 1. For a convergence criterion, we may compare the norm of the change in (B, Ω) or the resulting covariance matrix or the value of $\ell_{G,Y}$ to a given tolerance.

Because cases (ii) and (iii) of Theorem 1 allow for nonunique or nonexistent solutions to block update problems, a remaining concern is whether the BCD algorithm may fail to be well defined. We address this problem in Section 4, where

Algorithm 1 Block-coordinate descent

Require: Y, $\Omega^{(0)}$ and $B^{(0)}$

repeat
  for $i \in V$ **do**
    Fix $\Omega_{-i,-i}$ and B_{-i}, and compute residuals ε_{-i} and pseudo-variables $Z_{\mathrm{sib}(i)}$
    Compute $c_{i,0}$ and $c_{i,\mathrm{pa}(i)}$ as in Lemma 2
    if $c_{i,\mathrm{pa}(i)} \neq 0$ **then**
      Set up problem (3.8) with $y = Y_i^T$, $X = (Y_{\mathrm{pa}(i)}^T, Z_{\mathrm{sib}(i)}^T)$, $c = (c_{i,\mathrm{pa}(i)}^T, 0)^T$ and $c_0 = c_{i,0}$
      Compute an orthogonal matrix Q_1 with $Q_1 c = (0, \ldots, 0, \|c\|)^T$
      Compute QR decomposition $Q_1 X = Q_2^T R$
      Extract submatrix $R_1 = R_{\{1,\ldots,m\}\times\{1,\ldots,m\}}$
      Compute intermediate constants $r = R_{mm}$, y_0^2, and $(Q_2 y)_j$ for $j = 1, \ldots, m$
      Compute α'' using (3.13) and (3.15)
      Compute $(\hat{B}_{i,\mathrm{pa}(i)}, \hat{\Omega}_{i,\mathrm{sib}(i)})^T = \alpha = Q_1^T R_1^{-1} \alpha''$
    else
      Compute $(\hat{B}_{i,\mathrm{pa}(i)}, \hat{\Omega}_{i,\mathrm{sib}(i)})$ by minimizing sum of squares in numerator of (3.7)
    end if
    Compute $\hat{\omega}_{ii.-i}$ using (3.5)
    Update B_i and $\Omega_{i,-i} = \Omega_{-i,i}^T$ using $\hat{B}_{i,\mathrm{pa}(i)}$ and $\hat{\Omega}_{i,\mathrm{sib}(i)}$, respectively
    Update Ω by setting $\omega_{ii} = \hat{\omega}_{ii.-i} + \Omega_{i,-i}\Omega_{-i,-i}^{-1}\Omega_{-i,i}$
  end for
until Convergence criterion is met

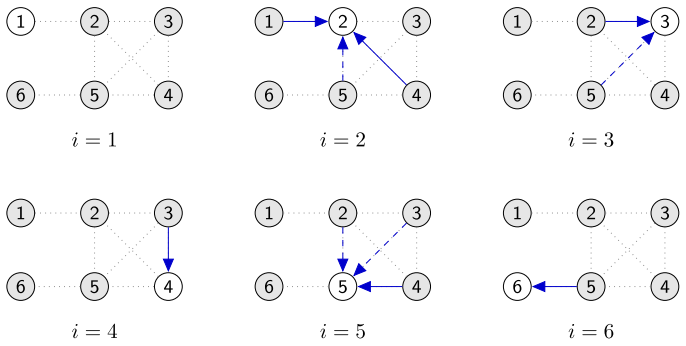

FIG. 3. *Illustration of the update steps for the BCD algorithm for each node. At each step, the edges corresponding to fixed parameters have been replaced with dotted edges. Arrowheads into nodes other than i have been removed. Hence, solid and dashed directed edges into i respectively represent directed and bi-directed edges with an arrowhead at i. Each remaining arrowhead signifies the relevant parameter to update during this step.*

we characterize the mixed graphs for which BCD updates are unique and feasible. This characterization treats generic data Y and generically chosen starting values for (B, Ω). As discussed in Section 4.4, graphs for which the BCD algorithm is not generically well defined yield nonidentifiable models. Identifiability is not necessary, however, for the BCD algorithm to be generically well defined. Furthermore, nonuniqueness of block update solutions could be addressed as outlined in Remark 2.

EXAMPLE 2. We illustrate the BCD algorithm for the graph from Figure 2, visiting the nodes in the order of their labels from 1 to 6. Since the graph is simple (i.e., without bows), the theory from Section 4.2 shows that all updates depicted in Figure 3 are well defined.

Beginning with node $i = 1$, we fix all but the first row of B and the first row and column of Ω. In graphical terms, we fix the parameters that correspond to edges that do not have an arrowhead at node 1. Now, there are no arrowheads at node 1, and all entries in the first row of B and all off-diagonal entries in the first row and column of Ω are zero. Consequently, the algorithm merely updates the variance ω_{11}. The update simply sets $\omega_{11} = S_{11}$, the sample variance for variable 1. This update is the same in later iterations, that is, node 1 can be skipped in subsequent iterations.

For $i = 2$, three edges have arrowheads at node 2, with corresponding parameters β_{21}, β_{24} and ω_{25}. The directed edge $4 \to 2$ is contained in a cycle of the graph. Its associated parameter, β_{24}, has coefficient $-\beta_{32}\beta_{43}$ in $\det(I - B)$. Thus, unless β_{32} or β_{43} is zero, $c_{2,\mathrm{pa}(2)} \neq 0$ and the more involved update from lines 6–11 in Algorithm 1 applies. If β_{32} or β_{43} is fixed to zero during this first iteration of the algorithm (i.e., one or both were initialized to zero), then the first update for $i = 2$

is a least squares problem but subsequent updates would almost surely require the more involved update.

Nodes 3 and 4 each have one arrowhead corresponding to a directed edge contained in a cycle of the graph. Hence, the updates for $i \in \{3, 4\}$ proceed analogously to the update step $i = 2$. For $i = 3$, we update the parameters β_{32}, ω_{35}, and ω_{33}. For $i = 4$, we update the parameters β_{43} and ω_{44}.

For $i = 5$, there are three arrowheads at node 5 corresponding to parameters β_{54}, ω_{25}, and ω_{35}. Observe that $4 \to 5$ is the only directed edge into node 5 and is not contained in a cycle. Hence $c_{5,\mathrm{pa}(5)} = 0$, and we proceed with the least squares update in line 13 of Algorithm 1. This least squares computation may change from one iteration of the algorithm to the next.

For $i = 6$, the only arrowhead corresponds to the directed edge $5 \to 6$ with associated parameter β_{65}. This directed edge is not involved in a cycle, so we estimate the parameter via a least squares regression and then solve for ω_{66}. This update remains the same throughout all iterations of the algorithm and only needs to be performed once.

4. Properties of the block-coordinate descent algorithm.

4.1. *Convergence properties*. Since the BCD algorithm performs partial maximizations, the value of the log-likelihood function $\ell_{G,Y}$ is nondecreasing throughout the iterations. When initialized generically, the algorithm finds a positive definite covariance matrix at every update. The update steps preserve the structural zeros of the matrices B and Ω, and $I - B$ remains invertible. Hence, the algorithm constructs a sequence in $\mathbf{B}(G) \times \mathbf{\Omega}(G)$.

Every accumulation point $(B^\star, \Omega^\star)$ of the sequence constructed by the algorithm is a critical point of the likelihood function and either a local maximum or a saddle point. A local maximum can be certified by checking negative definiteness of the Hessian of $\ell_{G,Y}$. However, as "always" in general nonlinear optimization there is no guarantee that a global maximum is found. Indeed, even for seemingly simple mixed graphs, the likelihood function can be multimodal [Drton and Richardson (2004)]. In practice, one may wish to run the algorithm from several different initial values. A strength of the BCD algorithm is that for nodes whose incoming directed edges are not contained in any cycle of G and that are not incident to any bi-directed edges, the update of $B_{i,\mathrm{pa}(i)}$ and ω_{ii} does not depend on the fixed pair $(B_{-i}, \Omega_{-i,-i})$, and thus needs to performed only once (in the first iteration). As we had noted, this happens for nodes 1 and 6 of the example discussed in Section 3.4. Hence, we may check for nodes of this type and exclude them from subsequent iterations after the first iteration of the algorithm. We also update these nodes before the set of nodes that require multiple update iterations.

4.2. *Existence and uniqueness of optima in block updates.* The BCD algorithm is well defined if each block update problem has a unique solution that is feasible, where feasibility refers to the new matrix Ω being positive definite. When updating at node i, the positive definiteness of Ω is equivalent to $\omega_{ii.-i} > 0$. Since the latter conditional variance is set via (3.5), feasibility of a block update solution $(\Omega_{i,\mathrm{sib}(i)}, B_{i,\mathrm{pa}(i)})$ corresponds to $\|Y_i - B_{i,\mathrm{pa}(i)}Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)}Z_{\mathrm{sib}(i)}\|$ being positive.

If the underlying graph is acyclic, then the update at node i solves a least squares problem that has a unique solution if and only if the $|\mathrm{pa}(i)| + |\mathrm{sib}(i)|$ vectors in the rows of $Y_{\mathrm{pa}(i)}$ and $Z_{\mathrm{sib}(i)}$ form a linearly independent set in $\mathbb{R}^N$. Moreover, the update yields a positive value of $\omega_{ii.-i}$ if and only if Y_i is not in the linear span of the rows of $Y_{\mathrm{pa}(i)}$ and $Z_{\mathrm{sib}(i)}$. We conclude that, in the acyclic case, the block update admits a unique and feasible solution if and only if the following condition is met:

$(\mathrm{A1})_i$ The matrix $\binom{Z_{\mathrm{sib}(i)}}{Y_{\mathrm{pa}(i)\cup\{i\}}} \in \mathbb{R}^{(|\mathrm{sib}(i)|+|\mathrm{pa}(i)|+1)\times N}$ has linearly independent rows.

As we show in Theorem 2 below, if the underlying graph is not acyclic, then a further condition is needed:

$(\mathrm{A2})_i$ The inequality $c_{i,0} + \hat{B}_{i,\mathrm{pa}(i)}c_{i,\mathrm{pa}(i)} \neq 0$ holds for $\hat{B}_{i,\mathrm{pa}(i)} = [Y_i X_i^T (X_i X_i^T)^{-1}]_{\mathrm{pa}(i)}$ and $X_i = \begin{pmatrix} Z_{\mathrm{sib}(i)} \\ Y_{\mathrm{pa}(i)} \end{pmatrix}$.

Note that the acyclic case has $c_{i,0} = 1$ and $c_{i,\mathrm{pa}(i)} = 0$, so condition $(\mathrm{A2})_i$ is void.

EXAMPLE 3. Let the graph $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be a two-cycle, so $V = \{1, 2\}$, $E_{\rightarrow} = \{1 \rightarrow 2, 2 \rightarrow 1\}$ and $E_{\leftrightarrow} = \varnothing$. Consider the update for node $i = 2$. With $\mathrm{pa}(2) = 1$, we have $c_{2,\mathrm{pa}(2)} = -\beta_{12}$ and $c_{2,0} = 1$. Since $\mathrm{sib}(2) = \varnothing$, the block update amounts to solving

$$\min_{\beta_{21}\in\mathbb{R}} \frac{\|Y_2 - \beta_{21}Y_1\|^2}{(1 - \beta_{12}\beta_{21})^2}$$

for fixed β_{12}. Condition $(\mathrm{A1})_i$ holds for $i = 2$ when the data vectors Y_1 and and Y_2 are linearly independent. We are then in case (i) or (iii) of Theorem 1. Hence, the solution either exists uniquely or does not exist. It fails to exist when

$$1 - \beta_{12}\frac{\langle Y_2, Y_1\rangle}{\|Y_1\|^2} = 0,$$

that is, when $(\mathrm{A2})_i$ fails for $i = 2$.

THEOREM 2. *Let $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be any mixed graph, and let $Y \in \mathbb{R}^{V\times N}$ be a data matrix of full rank $|V| \leq N$. Let $i \in V$ be any node. Then the function g_i*

from (3.7) *has a unique minimizer* $(\Omega_{i,\mathrm{sib}(i)}, B_{i,\mathrm{pa}(i)})$ *with*

$$\|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)}\| > 0$$

if and only if conditions $(\mathrm{A1})_i$ *and* $(\mathrm{A2})_i$ *hold.*

PROOF. ($\Longleftarrow$) When $(\mathrm{A1})_i$ holds, Theorem 1 applies to the minimization of g_i because the matrix X defined in (3.9) has full rank. Condition $(\mathrm{A2})_i$ ensures we are in case (i) of the theorem. Hence, g_i has a unique minimizer $(\Omega_{i,\mathrm{sib}(i)}, B_{i,\mathrm{pa}(i)})$. According to $(\mathrm{A1})_i$, Y_i^T is not in the span of X. Thus, $Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)} \neq 0$.

($\Longrightarrow$) First, suppose $(\mathrm{A1})_i$ holds but $(\mathrm{A2})_i$ fails. Then Theorem 1 applies in either case (ii) or (iii). Hence, the minimizer of g_i is either not unique or does not exist.

Second, suppose condition $(\mathrm{A1})_i$ fails because $X = (Z_{\mathrm{sib}(i)}^T, Y_{\mathrm{pa}(i)}^T)$ is not of full rank. Let $\eta \in \mathbb{R}^{|\mathrm{sib}(i)|+|\mathrm{pa}(i)|}$ be any nonzero vector in the kernel of X. Let $c = (0, c_{i,\mathrm{pa}(i)}^T)^T$. With the orthogonality from Lemma 4, we have $X\alpha = X(\alpha + \eta)$ and $c^T\alpha = c^T(\alpha + \eta)$ for any $\alpha \in \mathbb{R}^{|\mathrm{sib}(i)|+|\mathrm{pa}(i)|}$. Consequently, g_i does not have a unique minimizer.

Third, suppose that $X = (Z_{\mathrm{sib}(i)}^T, Y_{\mathrm{pa}(i)}^T)$ has full rank but $(\mathrm{A1})_i$ still fails. Then $y = Y_i^T$ is in the column span of X so that Theorem 1 applies with the quantity y_0^2 zero. We are thus in either case (i) or case (ii) of the theorem. In case (ii), the minimizer is not unique. This leaves us with case (i), in which $y_0^2 = 0$ implies that g_i is uniquely minimized by the least squares vector $\hat{\alpha}$, that is, the minimizer of $\alpha \mapsto \|y - X\alpha\|^2$. Since $y = Y_i^T$ is in the span of X, we have $\|y - X\alpha\|^2 = 0$, which translates into $Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sib}(i)} Z_{\mathrm{sib}(i)} = 0$. We conclude that g_i has a unique and feasible minimizer only if $(\mathrm{A1})_i$ and $(\mathrm{A2})_i$ hold. □

EXAMPLE 4. Let $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be the graph with vertex set $V = \{1, 2\}$, and edge sets $E_{\rightarrow} = \{1 \rightarrow 2, 2 \rightarrow 1\}$ and $E_{\leftrightarrow} = \{1 \leftrightarrow 2\}$. Note that the model $\mathbf{N}(G)$ comprises all centered bivariate normal distributions. Therefore, the log-likelihood function $\ell_{G,Y}$ achieves its maximum for any data matrix $Y \in \mathbb{R}^{2\times N}$ of rank 2.

The two block updates in this example are symmetric, so consider the update for $i = 1$ only. Fix any two values of $\beta_{21} \in \mathbb{R}$ and $\omega_{22} > 0$. Then the map from $(\beta_{12}, \omega_{12}, \omega_{11})$ to the covariance matrix $(I - B)^{-1}\Omega(I - B)^{-T}$ is easily seen to have a Jacobian matrix of rank 2. Because the rank drops from 3 to 2, for each triple $(\beta_{12}, \omega_{12}, \omega_{11})$ there is a one-dimensional set of other triples that yield the same covariance matrix, and thus, the same value of the likelihood function. Due to this lack of blockwise identifiability, the block update cannot have a unique solution.

In this example, we have $\mathrm{sib}(1) = \mathrm{pa}(1) = \{2\}$ and $\det(I - B) = 1 - \beta_{12}\beta_{21}$, so that $c_{i,0} = 1$ and $c = (0, -\beta_{21})^T$. Moreover,

$$X^T = \begin{pmatrix} Z_{\mathrm{sib}(i)} \\ Y_{\mathrm{pa}(i)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\omega_{22}}(Y_2 - \beta_{21}Y_1) \\ Y_2 \end{pmatrix} = \begin{pmatrix} -\frac{\beta_{21}}{\omega_{22}} & \frac{1}{\omega_{22}} \\ 0 & 1 \end{pmatrix} Y.$$

If $\beta_{21} = 0$, then $(\mathrm{A1})_i$ fails for $i = 1$ because X is rank deficient. If $\beta_{21} \neq 0$, and $\mathrm{rank}(Y) = 2$, then $\mathrm{rank}(X) = 2$ and $y = Y_1^T$ is in the span of X, with

$$Y_1 = \begin{pmatrix} -\frac{\omega_{22}}{\beta_{21}} & \frac{1}{\beta_{21}} \end{pmatrix} X^T.$$

Consequently, $y_0^2 = 0$ and the least squares coefficients for the regression of y on X are $(-\omega_{22}/\beta_{21}, 1/\beta_{21})$. Then condition $(\mathrm{A2})_i$ fails for $i = 1$ because with $c_{1,\mathrm{pa}(1)} = -\beta_{21}$ and least squares coefficient $\hat{B}_{1,\mathrm{pa}(1)} = 1/\beta_{21}$ we find that

$$c_{1,0} + \hat{B}_{1,\mathrm{pa}(1)} c_{1,\mathrm{pa}(1)} = 1 + \frac{1}{\beta_{21}}(-\beta_{21}) = 0.$$

REMARK 3. The findings from Example 4 generalize. Indeed, for any graph G, if Y has full rank and Y_i^T is in the span of $X = (Z_{\mathrm{sib}(i)}^T, Y_{\mathrm{pa}(i)}^T)$, then one can show that $(\mathrm{A2})_i$ fails, and thus, the block update has infinitely many solutions; see the Supplementary Material [Drton, Fox and Wang (2018)].

4.3. *Well-defined BCD iterations.* Although Theorem 2 characterizes the existence of a unique feasible solution for a particular block update, it does not yet clarify when its conditions $(\mathrm{A1})_i$ and $(\mathrm{A2})_i$ hold throughout all iterations of the BCD algorithm. In practice, there is freedom in choosing the starting value $(B_0, \Omega_0) \in \mathbf{B}(G) \times \mathbf{\Omega}(G)$ and, in particular, we may choose it randomly to alleviate problems of having the triple (Y, B_0, Ω_0) in undesired special position; recall Example 3. Since our models concern a continuously distributed data matrix $Y \in \mathbb{R}^{V \times N}$, the natural problem is to characterize the graphs G such that any finite number of BCD iterations are well defined for generic triples (Y, B_0, Ω_0). As before, our treatment assumes $N \geq |V|$.

We begin by studying condition $(\mathrm{A1})_i$. Let $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be a mixed graph. Let π be a path in G, and let $i_1, \ldots, i_k$ be the not necessarily distinct vertices on π. Then π is a *half-collider path* if either all edges on π are bi-directed, or the first edges is $i_1 \rightarrow i_2$ and all other edges are bi-directed. Both a single edge $i_1 \rightarrow i_2$ and an empty path comprising only node i_1 are half-collider paths. The *bi-directed portion* of a half-collider path π is the set of nodes that are incident to a bi-directed edge on π. In other words, if π starts with $i_1 \rightarrow i_2$, then its bi-directed portion is $\{i_2, \ldots, i_k\}$. If π does not contain a directed edge, then its bi-directed portion is the set of all of its nodes $\{i_1, \ldots, i_k\}$. Valid half-collider paths are shown in Figure 4.

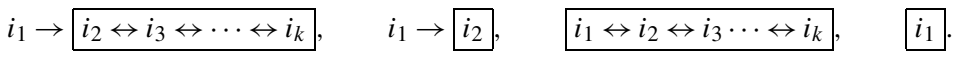

FIG. 4. *Four half-collider paths with boxes around their bi-directed portions.*

We note that half-collider paths are dual to the half-treks of Foygel, Draisma and Drton (2012). A half-trek is a path whose first edge is either directed or bi-directed, and whose remaining edges are directed.

Let $S_b, S_e \subset V$ be two sets of nodes. A collection of paths $\pi^1, \ldots, \pi^s$ is a *system of half-collider paths* from S_b to S_e if $|S_b| = |S_e| = s$, each π^l is a half-collider path from a node in S_b to a node in S_e, every node in S_b is the first node on some π^l, and every node in S_e is the last node on some π^l.

PROPOSITION 2. *Let $G = (V, E_{\to}, E_{\leftrightarrow})$ be a mixed graph, and let $i \in V$. Then the following two statements are equivalent*:

(a) *Condition* (A1)$_i$ *holds for generic triples* $(Y, B, \Omega) \in \mathbb{R}^{V \times N} \times \mathbf{B}(G) \times \mathbf{\Omega}(G)$.

(b) *The induced subgraph G_{-i} contains a system of half-collider paths from some subset $S_b(i) \subseteq V \setminus (\mathrm{pa}(i) \cup \{i\})$ to $S_e(i) = \mathrm{sib}(i)$ such that the bi-directed portions are pairwise disjoint.*

The proof is deferred to the Supplementary Material [Drton, Fox and Wang (2018)]; it merely requires Y to be of full rank and (B, Ω) to be chosen from a set of generic points that is independent of Y.

EXAMPLE 5. Suppose a graph with vertex set $V = \{1, \ldots, 6\}$ contains the paths

$$1 \to 3 \leftrightarrow 4 \leftrightarrow 5 \quad \text{and} \quad 2 \leftrightarrow 1 \leftrightarrow 6.$$

These form a system of half-collider paths from $\{1, 2\}$ to $\{5, 6\}$. The system is not vertex disjoint as node 1 appears on both paths. However, the bi-directed portions $\{3, 4, 5\}$ and $\{1, 2, 6\}$ are disjoint.

Next, we turn to condition (A2)$_i$ and show that in generic cases it does not impose any additional restriction.

PROPOSITION 3. *Suppose the mixed graph G is such that* (A1)$_i$ *holds for generic triples* $(Y, B, \Omega) \in \mathbb{R}^{V \times N} \times \mathbf{B}(G) \times \mathbf{\Omega}(G)$. *Then* (A2)$_i$ *holds for generic triples* (Y, B, Ω).

PROOF. The matrix X_i and the least squares vector $\hat{B}_{i,\mathrm{pa}(i)}$ in condition (A2)$_i$ are rational functions of the triple (Y, B, Ω). Hence, there is a polynomial

$f(Y, B, \Omega)$ such that $(A2)_i$ fails only if f vanishes. A polynomial that is not the zero polynomial has a zero set that is of reduced dimension and of measure zero [Okamoto (1973), Lemma 1]. Therefore, it suffices to show that $(A2)_i$ holds for a single choice of (Y, B, Ω).

By assumption, we may pick $B \in \mathbf{B}(G)$ and $\Omega \in \mathbf{\Omega}(G)$ such that $(A1)_i$ holds for any full rank Y. Take Y such that $(I - B)^{-1}\Omega(I - B)^{-T} = \frac{1}{N}YY^T$. When Y has full rank, the normal distribution with covariance matrix $\frac{1}{N}YY^T$ has maximal likelihood. Therefore, Ω and B maximize the log-likelihood function $\ell_{G,Y}$. Consider now the block update for node i. Because $(A1)_i$ holds, the matrix X_i has full rank and Y_i is not in the span of X_i. Hence, Theorem 1 applies with $y_0^2 > 0$. Since our special choice of (Y, B, Ω) guarantees the existence of an optimal solution, we must be in case (i) of the theorem. The inequality defining this case corresponds to $(A2)_i$. □

The following theorem gives a combinatorial characterization of the graphs for which the BCD algorithm is well defined. It readily follows from the above results, as we show in the Supplementary Material [Drton, Fox and Wang (2018)].

THEOREM 3. *For a mixed graph $G = (V, E_{\to}, E_{\leftrightarrow})$, the following two statements are equivalent*:

(a) *For all $i \in V$, the induced subgraph G_{-i} contains a system of half-collider paths from a set of nodes $S_b(i) \subseteq V \setminus (\mathrm{pa}(i) \cup \{i\})$ to $S_e(i) = \mathrm{sib}(i)$ such that the bi-directed portions are pairwise disjoint.*

(b) *For generic triples $(Y, B_0, \Omega_0) \in \mathbb{R}^{V \times N} \times \mathbf{B}(G) \times \mathbf{\Omega}(G)$, any finite number of iterations of the BCD algorithm for $\ell_{G,Y}$ have unique and feasible block updates when (B_0, Ω_0) is used as starting value.*

In Drton, Eichler and Richardson (2009), the focus was on bow-free acyclic graphs, where bow-free means that there do not exist two nodes i and j with both $i \to j$ and $i \leftrightarrow j$ in G. For such graphs, the BCD algorithm is easily seen to be well defined. More generally, by taking $S_b(i) = \mathrm{sib}(i)$ we obtain the following generalization to graphs that may contain directed cycles.

PROPOSITION 4. *If G is a simple mixed graph, that is, every pair of nodes is incident to at most one edge, then condition (a) in Theorem 3 holds.*

When the graph G is not simple, checking condition (a) from Theorem 3 is more involved. It can, however, be checked in polynomial time. As noted in condition (b), well-defined updates are only guaranteed for generic initializations. In particular, if β_{ij} is initialized to 0, this effectively removes the $j \to i$ edge from the graph and will cause the update to become ill defined if that edge is necessary for condition (a) to hold.

PROPOSITION 5. *For any mixed graph* $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$, *condition* (*a*) *in Theorem* 3 *can be checked in* $\mathcal{O}(|V|^5)$ *operations*.

The proof, which is deferred to the Supplementary Material [Drton, Fox and Wang (2018)], casts checking the condition as a network flow problem.

4.4. *Identifiability.* There is a close connection between well-defined block updates and parameter identifiability. Suppose the data matrix Y is such that the sample covariance is $S = \frac{1}{N} YY^T = (I - B)^{-1}\Omega(I - B)^{-T}$ for a pair $(B, \Omega) \in \mathbf{B}(G) \times \mathbf{\Omega}(G)$. Consider the block update of the ith row of B and ith row and column of Ω. Based on Theorem 1, if the update does not have a unique solution then there is an infinite set of solutions (B', Ω'). Each such solution (B', Ω') must have $(I - B')^{-1}\Omega'(I - B')^{-T}$ equal to S because S is the unique covariance matrix with maximum likelihood. Hence, there is an infinite set of parameters (B', Ω') that define the same normal distribution as (B, Ω).

COROLLARY 1. *If the graphical condition in statement* (*a*) *of Theorem* 3 *fails for the graph* G, *then the parameters of model* $\mathbf{N}(G)$ *are not identifiable*.

5. Simulation studies. In this section, we analyze the performance of our BCD algorithm in two contexts. First, we use it to compare the fit of two nested models (one of which is cyclic) for data on protein abundances. Second, we examine the problem of parameter estimation in a specified model. There we compare our algorithm on a number of simulated graphs against the fitting routine from the "sem" package in R [Fox (2006), R Development Core Team (2011)].

5.1. *Protein-signaling network.* Figure 2 in Sachs et al. (2005) presents a protein-signaling network involving 24 molecules. Abundance measurements are available for 11 of these. The remaining 13 are unobserved. For our illustration, we select two plausible mixed graphs over the 11 observed variables. The graphs differ only by the presence of a directed edge that induces a cycle and a bow; see Figure 5. The edge PIP2 $\rightarrow$ PIP3, which makes for the difference, is highlighted in red. Before proceeding to our analysis, we note that the results in Sachs et al. (2005) are based on discretized data and are thus not directly comparable to our computations.

We proceed by comparing the two candidate models via the likelihood ratio test. The data we consider consist of 11 simultaneously observed signaling molecules measured independently across $N = 853$ individual primary human immune system cells. Specifically, we consider the data from experimental condition CD3 + CD28 and center/rescale the data, ensuring that each variable has zero mean and variance one. Although the likelihood ratio test statistic is invariant to scale, the rescaling improves the conditioning of the sample covariance matrix which improves the performance of BCD.

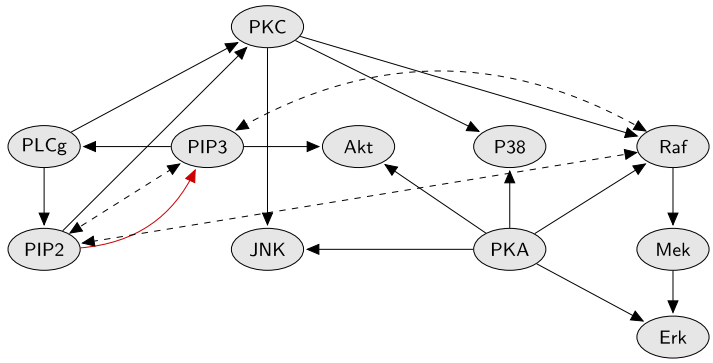

FIG. 5. *Plausible mixed graph for the protein-signaling network dataset. The relevant acyclic submodel can be formed by removing the red directed edge from PIP*2 *to PIP*3.

The corresponding likelihood ratio test statistic for the data is 0.075, and under the standard χ^2_1 asymptotic distribution for the null hypothesis, this corresponds to a p-value of 0.78. However, in the considered models it is not immediately clear whether a χ^2_1 approximation has (asymptotic) validity, as the models generally have a singular parameter space [Drton (2009)]. Therefore, we enlist subsampling as a guard against a possible nonstandard asymptotic distribution. Subsampling only requires the existence of a limiting distribution for the likelihood ratio statistic [Politis, Romano and Wolf (1999), Chapter 2.6]. This limiting distribution, while not necessarily chi-squared, is guaranteed to always exist [Drton (2009)]. Each random subsample consists of b observations where b is chosen large enough to approximate the true asymptotic distribution under the null, but small compared to $N = 853$ to still provide reasonable power under the alternative. We consider 5000 subsamples of sizes $b = 30$ and $b = 50$.

For each subsample, we first fit the sub-model corresponding to the mixed graph depicted in Figure 5 without the edge PIP2 $\to$ PIP3. For this procedure, we initialize the free entries of B using least squares regression estimates (i.e., fitting the model that ignores the error correlations). We then calculate the covariance between the regression residuals to estimate the nonzero elements of Ω. Although the sample covariance of the regression residuals is positive definite, the resulting matrix which also encodes the structural zeros may not be. To ensure that Ω is positive definite, we scale the off-diagonal elements so that the matrix is diagonally dominant. Specifically, for any row i where $\sum_{i \neq j} |\omega_{ij}| > \omega_{ii}$, we rescale the off-diagonal elements so that $\sum_{i \neq j} |\omega_{ij}| = 0.9 \times \omega_{ii}$ and set $\omega_{ji} = \omega_{ij}$. If a row is already diagonally dominant, we do not explicitly rescale the off-diagonal elements, but individual elements of the row might be modified to preserve symmetry after other rows have been rescaled. After the BCD algorithm converges to a stationary point in the sub-model, we take the fitted values $\hat{B}$ and $\hat{\Omega}$ to initialize the algorithm run on the model that includes the additional PIP2 $\to$ PIP3 edge.

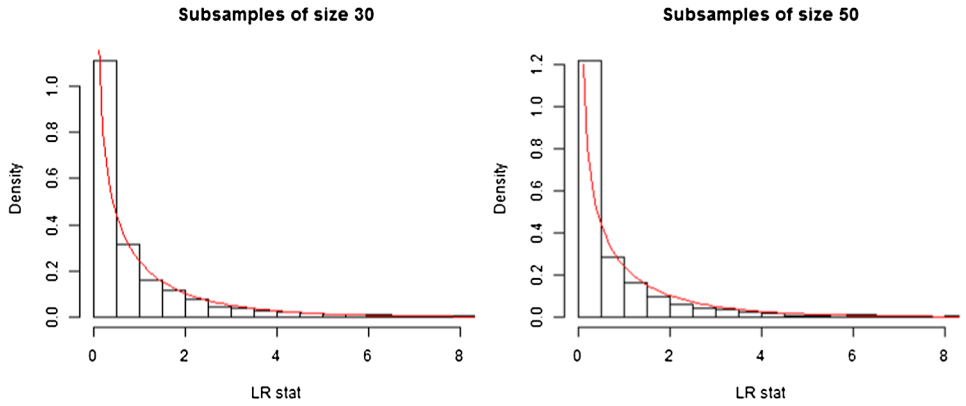

FIG. 6. *Histograms for the likelihood ratio test statistic for* 5000 *subsamples of size* 30 *and* 50, *respectively. The superimposed red line depicts the* χ_1^2 *density.*

We evaluate the likelihood function at each of the two maxima and formulate the corresponding likelihood ratio test statistic. The choice of $\hat{B}$ and $\hat{\Omega}$ as initial values for estimating the larger model guarantees that the test statistics are nonnegative.

Figure 6 shows histograms for the subsampled log-likelihood ratio statistics. The empirical distributions for $b = 30$ and $b = 50$ are seen to be similar to one another and also rather close to a χ_1^2 distribution. The observed test statistic for the full data has empirical p-value of 0.76 and 0.73 for $b = 30$ and $b = 50$, respectively. These p-values are slightly smaller than the p-value of 0.78 from χ_1^2 approximation. Altogether there is little evidence to reject the sub-model in favor of the more complicated cyclic model.

5.2. *Simulated data.* We now demonstrate how the BCD algorithm behaves on different types of mixed graphs. We consider the existing R package "sem" [Fox (2006)] as an alternative and compare the performance of these algorithms for maximum likelihood estimation on simulated data.

To simulate a mixed graph, we begin with the empty graph on V nodes. For $0 \le k \le V$, we add directed edges $1 \to 2 \to \cdots \to (k-1) \to k \to 1$, creating a directed cycle of length k. For all $p(p-1)/2 - (k-1)$ remaining pairs of nodes (i, j) with $i < j$, we generate independent uniform random variables $U_{ij} \sim U(0, 1)$. If $U_{ij} \le d$, we introduce the directed edge $i \to j$. Alternatively, if $d < U_{ij} \le b + d$, we introduce the bi-directed edge $i \leftrightarrow j$. If $U_{ij} > b + d$, there is no edge between i and j. After all edges have been determined, we randomly permute the node labels. This construction ensures that the resulting mixed graph G has the following properties:

(i) G has a unique cycle of length k;
(ii) G is simple when $k > 2$ and bow-free (for all k).

For this simulation, we use 24 different configurations of (V, N, k, d, b), where N is the sample size. We examine graphs of size $V = 10$ and $V = 20$ and $N = 3V/2$ and $N = 10V$ observations. In each of these 4 configurations, we consider 3 distinct choices of the maximum cycle length: $k = 0$, $V/5$, and $2V/5$. For each combination of (V, N, k), we let $d = 0.1$ and $d = 0.2$, fixing $b = d/2$ in each case. Note that in the case of $k = 0$, every generated graph will be acyclic and simple, the class of mixed graphs considered by Drton, Eichler and Richardson (2009).

In each simulation, we generate a random mixed graph G according to the procedure above. We then select a random distribution from the corresponding normal model $\mathbf{N}(G)$ by taking the covariance matrix to be $\Sigma = (I - B)^{-1}\Omega(I - B)^{-T}$ for $B \in \mathbf{B}(G)$ and $\Omega \in \mathbf{\Omega}(G)$ selected as follows. We set all free, off-diagonal entries of B and Ω to independent realizations from a $\mathcal{N}(0, 1)$ distribution. The diagonal entries of Ω are chosen as one more than the sum of the absolute values of the entries in the corresponding row of Ω plus a random draw from a χ^2_1 distribution. Hence, Ω is diagonally dominant and positive definite. The model $\mathbf{N}(G)$ is then fit to a sample of size N that is generated from the selected distribution. We use the routine "sem" and our BCD algorithm. We consider the BCD algorithm to have converged when $\frac{1}{|V|^2}\|\hat{\Sigma}^{(t-1)} - \hat{\Sigma}^{(t)}\| < 10^{-6}$ where $\|\cdot\|$ is the vector L_1 norm. The algorithm proceeds for a maximum of 5000 iterations, at which point divergence is assumed. The BCD algorithm is initialized using the procedure described in Section 5.1. The "sem" method is initialized by default using a modification of the procedure described by McDonald and Hartmann (1992).

Each row of Table 1 corresponds to 1000 simulations at a configuration of (V, N, k, d, b). We record how often each algorithm converges. The columns "both converge" and "both agree" report the number of simulations for which both algorithms converged, and the number of these simulations for which the resulting estimates were equal up to a small tolerance. For the routine "sem," which uses a generic "nlm" Newton optimizer, it is not uncommon that convergence occurs but yields estimates that are not positive definite. In these cases, we consider the algorithm to have not converged.

The last two columns show the average CPU running times (in milliseconds) over simulations for which both methods converged and agreed.[1] We caution that these times are not directly comparable, since "sem" computes a number of other quantities of interest in addition to the maximum likelihood estimate. However, the BCD algorithm is up to 6 times faster than "sem" in some instances. One potential reason is that when the graph is relatively sparse, many of the nodes may only require a single BCD update.

[1]The simulations were run on a laptop with a quad-core 2.4 Ghz processor.

TABLE 1

Data simulated from a random distribution in a randomly generated mixed graph model is fit to the model using BCD and the quasi-Newton method invoked by "sem." Each row summarizes 1000 *simulations. "Both agree" counts the cases with ML estimates equal up to small tolerance. Running time is average CPU time (in milliseconds) for the cases in which both algorithms converged and agreed*

				Convergence		Both converge	Both agree	Running time	
V	*N*	*k*	*d*	BCD	SEM			BCD	SEM
10	15	0	0.1	1000	991	991	932	3.8	24.1
10	15	0	0.2	1000	949	949	884	9.5	31.8
10	15	2	0.1	1000	479	479	456	10.7	28.7
10	15	2	0.2	1000	559	559	518	16.0	36.2
10	15	4	0.1	997	672	672	637	10.7	30.5
10	15	4	0.2	997	553	553	520	16.7	38.0
10	100	0	0.1	1000	996	996	985	6.5	30.9
10	100	0	0.2	1000	991	991	991	20.9	53.3
10	100	2	0.1	1000	517	517	517	40.1	48.0
10	100	2	0.2	1000	635	635	635	51.5	58.9
10	100	4	0.1	999	726	726	725	33.4	50.2
10	100	4	0.2	998	688	688	688	46.3	63.0
20	30	0	0.1	1000	989	989	971	54.0	324.7
20	30	0	0.2	1000	921	921	881	166.7	550.5
20	30	4	0.1	999	836	836	824	77.3	319.5
20	30	4	0.2	998	731	731	701	197.0	652.6
20	30	8	0.1	1000	709	709	696	97.0	342.2
20	30	8	0.2	999	534	534	505	237.5	766.3
20	200	0	0.1	1000	998	998	993	119.8	330.1
20	200	0	0.2	1000	983	983	958	299.0	585.4
20	200	4	0.1	1000	847	847	829	199.5	356.8
20	200	4	0.2	999	806	806	773	359.6	712.3
20	200	8	0.1	999	765	765	755	257.6	409.8
20	200	8	0.2	1000	659	659	630	471.7	851.4

6. Discussion. This work extends the RICF algorithm of Drton, Eichler and Richardson (2009) to cyclic models. The RICF algorithm and its BCD extension iteratively perform partial maximizations of the likelihood function via joint updates to the parameter matrices B and Ω. Each update problem admits a unique solution. Like its predecessor, the generalized algorithm is guaranteed to produce feasible positive definite covariance matrices after every iteration. Moreover, any accumulation point of the sequence of estimated covariance matrices is necessarily either a local maximum or a saddle point of the likelihood function.

Despite its desirable properties, the algorithm is not without limitations. As with any iterative maximization procedure, there is no guarantee that convergence of the algorithm is to a global maximum, due to possible multi-modality of the likeli-

hood function. In addition, for certain models the algorithm may be ill defined due to collinearity of the covariates and pseudo-covariates in our update step. However, we show that the models for which this occurs are nonidentifiable. Moreover, we give necessary and sufficient graphical conditions for generically well-defined updates, which were not previously known for the acyclic case.

In some of our simulated examples the BCD algorithm, which does not use any overall second-order information, needed many iterations to meet a convergence criterion. It is possible that in those cases a hybrid method that also considers quasi-Newton steps would converge more quickly. Nevertheless, our numerical experiments show that the BCD algorithm is competitive in terms of computation time with the generic optimization tools as used in the R package "sem" all the while alleviating convergence problems.

SUPPLEMENTARY MATERIAL

Proofs of claims (DOI: 10.1214/17-AOS1602SUPP; .pdf). The supplement provides proofs for claims made in Sections 2, 3 and 4. Specifically, we verify the form of $\det(I - B)$ as claimed in Lemma 1 and derive the likelihood equations with respect to Ω and B. We also verify the claims in Lemmas 4 and 5 which are required for the BCD algorithm described in the constructive proof of Theorem 1. Finally, we verify the claims in Section 4 which characterize graphs for which the BCD algorithm is well defined when initialized generically. In particular, we give a graphical condition and show that it can be checked in time which is a polynomial of the considered variables.

REFERENCES

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

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025

CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94** 199–216. MR2307904

COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

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

DRTON, M., EICHLER, M. and RICHARDSON, T. S. (2009). Computing maximum likelihood estimates in recursive linear models with correlated errors. *J. Mach. Learn. Res.* **10** 2329–2348. MR2563984

DRTON, M., FOX, C. and WANG, Y. S. (2018). Supplement to "Computation of maximum likelihood estimates in cyclic structural equation models." DOI:10.1214/17-AOS1602SUPP.

DRTON, M. and MAATHUIS, M. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application* **4** 365–393.

DRTON, M. and RICHARDSON, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91** 383–392. MR2081308

DRTON, M., STURMFELS, B. and SULLIVANT, S. (2009). *Lectures on Algebraic Statistics*. *Oberwolfach Seminars* **39**. Birkhäuser, Basel. MR2723140

FOX, J. (2006). Structural equation modeling with the sem package in R. *Struct. Equ. Model.* **13** 465–486. MR2240110

FOX, C. (2014). Interpretation and inference of linear structural equation models. Ph.D. thesis, Univ. Chicago.

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., et al. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature* **529** 390–393.

HARARY, F. (1962). The determinant of the adjacency matrix of a graph. *SIAM Rev.* **4** 202–210. MR0144330

HOYLE, R. H., ed. (2012). *Handbook of Structural Equation Modeling*. Guilford Press, New York.

KLINE, R. B. (2015). *Principles and Practice of Structural Equation Modeling*, 4th ed. Guilford Press, New York.

LACERDA, G., SPIRTES, P., RAMSEY, J. and HOYER, P. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-08) 366–374. AUAI Press, Corvallis, OR.

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

MCDONALD, R. P. and HARTMANN, W. M. (1992). A procedure for obtaining initial values of parameters in the RAM model. *Multivar. Behav. Res.* **27** 57–76.

MOOIJ, J. M. and HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the* 29*th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-13) (A. Nicholson and P. Smyth, eds.) 431–439. AUAI Press, Corvallis, OR.

NARAYANAN, A. (2012). A review of eight software packages for structural equation modeling. *Amer. Statist.* **66** 129–138.

NOWZOHOUR, C., MAATHUIS, M. and BÜHLMANN, P. (2015). Structure learning with bow-free acyclic path diagrams. Available at: arxiv:1508.01717.

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

PARK, G. and RASKUTTI, G. (2016). Identifiability assumptions and algorithm for directed graphical models with feedback. Available at: arxiv:1602.04418.

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

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. MR1707286

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RICHARDSON, T. (1996). A discovery algorithm for directed cyclic graphs. In *Uncertainty in Artificial Intelligence* (*Portland*, *OR*, 1996) 454–461. Morgan Kaufmann, San Francisco, CA. MR1617227

RICHARDSON, T. (1997). A characterization of Markov equivalence for directed cyclic graphs. *Internat. J. Approx. Reason.* **17** 107–162. MR1462712

ROSSEEL, Y. (2012). lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48** 1–36.

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. and NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529.

SILVA, R. (2013). A MCMC approach for learning the structure of Gaussian acyclic directed mixed graphs. In *Statistical Models for Data Analysis* (P. Giudici, S. Ingrassia and M. Vichi, eds.) 343–351. Springer, New York.

SPIRTES, P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence*: *Proceedings of the* 11*th Conference* (P. Besnard and S. Hanks, eds.) 491–498. Morgan Kaufmann, San Francisco, CA.

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

STATACORP (2013). STATA structural equation modeling reference manual. StataCorp LP, College Station, TX, Release 13.

STEIGER, J. H. (2001). Driving fast in reverse. *J. Amer. Statist. Assoc.* **96** 331–338.

TRIANTAFILLOU, S. and TSAMARDINOS, I. (2016). Score-based vs constraint-based causal learning in the presence of confounders. In *UAI* 2016 *Workshop on Causation*: *Foundation to Application* (F. Eberhardt, E. Bareinboim, M. Maathuis, J. Mooij and R. Silva, eds.). *CEUR Workshop Proceedings* **1792** 59–67.

WERMUTH, N. (2011). Probability distributions with summary graph structure. *Bernoulli* **17** 845–879. MR2817608

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
Y. S. WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195-4322
USA
E-MAIL: md5@uw.edu
ysamwang@uw.edu

C. FOX
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: chrisfox.galton@gmail.com