Journal of Machine Learning Research 10 (2009) 2329-2348 Submitted 3/08; Revised 1/09; Published 10/09

Computing Maximum Likelihood Estimates in Recursive Linear Models with Correlated Errors

Mathias Drton DRTON@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Michael Eichler M.EICHLER@MAASTRICHTUNIVERSITY.NL
Department of Quantitative Economics
Maastricht University
P.O. Box 616,
6200 MD Maastricht
The Netherlands

Thomas S. Richardson THOMASR@UW.EDU
Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322, USA

Editor: Michael I. Jordan

Abstract

In recursive linear models, the multivariate normal joint distribution of all variables exhibits a dependence structure induced by a recursive (or acyclic) system of linear structural equations. These linear models have a long tradition and appear in seemingly unrelated regressions, structural equation modelling, and approaches to causal inference. They are also related to Gaussian graphical models via a classical representation known as a path diagram. Despite the models' long history, a number of problems remain open. In this paper, we address the problem of computing maximum likelihood estimates in the subclass of 'bow-free' recursive linear models. The term 'bow-free' refers to the condition that the errors for variables i and j be uncorrelated if variable i occurs in the structural equation for variable j. We introduce a new algorithm, termed Residual Iterative Conditional Fitting (RICF), that can be implemented using only least squares computations. In contrast to existing algorithms, RICF has clear convergence properties and yields exact maximum likelihood estimates after the first iteration whenever the MLE is available in closed form.

Keywords: linear regression, maximum likelihood estimation, path diagram, structural equation model, recursive semi-Markov model, residual iterative conditional fitting

1. Introduction

A system of linear structural equations determines a linear model for a set of variables by dictating that, up to a random error term, each variable is equal to a linear combination of some of the remaining variables. Traditionally the errors are assumed to have a centered joint multivariate normal distribution. Presenting a formalism for simultaneously representing causal and statistical hypothe-

ses (Pearl, 2000; Spirtes et al., 2000), these normal linear models, which are also called *structural equation models*, are widely used in the social sciences (Bollen, 1989) and many other contexts.

In seminal work, Wright (1921, 1934) introduced *path diagrams*, which are useful graphical representations of structural equations. A path diagram is a graph with one vertex for each variable and directed and/or bi-directed edges. A directed edge $i \to j$ indicates that variable i appears as covariate in the equation for variable j. The directed edges are thus in correspondence with the *path coefficients*, that is, the coefficients appearing in the linear structural equations. A bi-directed edge $i \leftrightarrow j$ indicates correlation between the errors in the equations for variables i and j. Graphs of this kind are also considered by Shpitser and Pearl (2006), who refer to them as recursive semi-Markovian causal models.

1.1 A Motivating Example

We motivate the normal linear models analyzed here with the following example, which is adapted from a more complex longitudinal study considered in Robins (2008).

Consider a two-phase sequential intervention study examining the effect of exercise and diet on diabetes. In the first phase patients are randomly assigned to a number of hours of exercise per week (Ex) drawn from a log-normal distribution. At the end of this phase blood pressure (BP) levels are measured. In the second phase patients are randomly assigned to a strict calorie controlled diet that produces a change in body-mass index (ΔBMI). The assigned change in BMI, though still randomized, is drawn, by design, from a normal distribution with mean depending linearly on $X = \log(\text{Ex})$ and BP. The dependence here is due to practical and ethical considerations. Finally at the end of the second phase, triglyceride levels (Y) indicating diabetic status are measured.

A question of interest is whether or not there is an effect of X on the outcome Y that is not mediated through the dependence of ΔBMI on X and BP. In other words, if there had been no ethical or practical restrictions, and the assignment (ΔBMI) in the second phase was completely randomized and thus independent of BP and X, would there still be any dependence between X and Y? Note that due to underlying confounding factors such as life history and genetic background, we would expect to observe dependence between BP and Y even if the null hypothesis of no effect of X on Y was true and the second treatment (ΔBMI) was completely randomized.

Our model consists of two pieces. First, the design of the study dictates that

$$X = \alpha_0 + \varepsilon_X, \tag{1}$$

$$\Delta\text{BMI} = \gamma_0 + \gamma_1 X + \gamma_2\,\text{BP} + \varepsilon_{\Delta\text{BMI}}, \tag{2}$$

where $\varepsilon_X \sim \mathcal{N}(0,\sigma_X^2)$ and $\varepsilon_{\Delta\text{BMI}} \sim \mathcal{N}(0,\sigma_{\Delta\text{BMI}}^2)$ are independent. This assignment model is complemented by a model describing how BP and Y respond to the prior treatments:

$$\text{BP} = \beta_0 + \beta_1 X + \varepsilon_{\text{BP}}, \tag{3}$$

$$Y = \delta_0 + \delta_1 X + \delta_2\,\Delta\text{BMI} + \varepsilon_Y, \tag{4}$$

where $(\varepsilon_{\text{BP}},\varepsilon_Y)^t$ are centered bivariate normal and independent of ε_X and $\varepsilon_{\Delta\text{BMI}}$. We denote the variances of ε_{BP} and ε_Y by σ_{BP}^2 and σ_Y^2, respectively, and write $\sigma_{\text{BP},Y}$ for the possibly non-zero covariance of ε_{BP} and ε_Y. Figure 1 shows the path diagram for this structural equation model.

Equations (1), (2) and (3) simply specify conditional expectations that can be estimated in regressions. However, this is not the case in general with (4). Instead,

$$\text{E}\left[Y \mid X, \Delta\text{BMI}\right] = \overline{\delta}_0 + \overline{\delta}_1 X + \overline{\delta}_2\,\Delta\text{BMI}$$

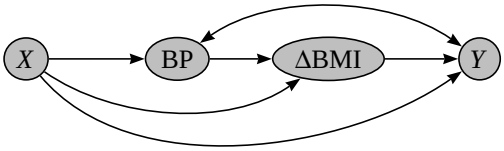

Figure 1: Path diagram illustrating a two-phase trial with two treatments (X and ΔBMI) and two responses (BP and Y). The treatment X is randomly assigned, and ΔBMI is randomized conditional on BP and X. The bi-directed edge indicates possible dependence due to unmeasured factors (genetic or environmental).

with

$$\bar{\delta}_1 = \delta_1 - \frac{\gamma_2 \sigma_{\mathrm{BP},Y}(\beta_1\gamma_2 + \gamma_1)}{\gamma_2^2\sigma_{\mathrm{BP}}^2 + \sigma_{\Delta\mathrm{BMI}}^2},$$

$$\bar{\delta}_2 = \delta_2 + \frac{\gamma_2 \sigma_{\mathrm{BP},Y}}{\gamma_2^2\sigma_{\mathrm{BP}}^2 + \sigma_{\Delta\mathrm{BMI}}^2},$$

and $\bar{\delta}_0 = \delta_0 + (\delta_1 - \bar{\delta}_1)\,\mathrm{E}[X] + (\delta_2 - \bar{\delta}_2)\,\mathrm{E}[\Delta\mathrm{BMI}]$. We see that δ_1 and δ_2 would have an interpretation as regression coefficients if: (i) the assignment of ΔBMI did not depend on BP (i.e., $\gamma_2 = 0$) and thus both treatments were completely randomized, or (ii) there were no dependence between ε_Y and $\varepsilon_{\mathrm{BP}}$ (i.e., $\sigma_{\mathrm{BP},Y} = 0$). Similarly, in $\mathrm{E}[Y \mid X, \mathrm{BP}, \Delta\mathrm{BMI}]$, the coefficient of ΔBMI is equal to δ_2 but the coefficient for X is $\delta_1 - \beta_1\sigma_{\mathrm{BP},Y}/\sigma_{\mathrm{BP}}^2$.

In this paper we consider likelihood-based methods for fitting a large class of structural equation models that includes the one given by (1)-(4) and can be used for consistent estimation of parameters such as δ_1. For alternative semi-parametric methods, see Robins (1999) and Gill and Robins (2001).

1.2 Challenges in Structural Equation Modelling

A number of mathematical and statistical problems arise in the normal linear models associated with systems of structural equations:

1. Different path diagrams may induce the same statistical model, that is, family of multivariate normal distributions. Such *model equivalence* occurs, for example, for the two path diagrams $1 \to 2$ and $1 \leftarrow 2$, which differ substantively by the direction of the cause-effect relationship. The two associated statistical models, however, are identical, both allowing for correlation between the two variables.

2. In many important special cases the path coefficient associated with a directed edge $i \to j$ has a population interpretation as a regression coefficient in a regression of j on a set of variables including i. However, as seen already in §1.1, this interpretation is not valid in general.

3. The parameters of the model may not be identifiable, so two different sets of parameter values may lead to the same population distribution; for an early review of this problem see Fisher (1966).

4. The set of parameterized covariance matrices may contain 'singularities' at which it cannot be approximated locally by a linear space. At 'singular' points, χ^2 and normal approximation to the distribution of likelihood ratio tests and maximum likelihood estimators (MLE) may not be valid; see, for instance, Drton (2009).

5. Iterative procedures are typically required for maximization of the likelihood function, which for some models can be multimodal (Drton and Richardson, 2004). Such multimodality typically occurs in small samples or under model misspecification.

The problems listed may arise in models without unobserved variables and become only more acute in latent variable models. They are challenging in full generality, but significant progress has been made in special cases such as recursive linear models with uncorrelated errors, which are also known as directed acyclic graph (DAG) models or 'Bayesian' networks (Lauritzen, 1996; Pearl, 1988). A normal DAG model is equivalent to a series of linear regressions, is always identified and has standard asymptotics. Under simple sample size conditions, the MLE exists almost surely and is a rational function of the data. Graphical modelling theory also solves problem 1 by characterizing all DAGs that induce the same statistical model (Andersson et al., 1997). For more recent progress on the general equivalence problem see Ali et al. (2009, 2005) and Zhang and Spirtes (2005).

1.3 Contribution of This Work

The requirement of uncorrelated errors may be overly restrictive in many settings. While arbitrary correlation patterns over the errors may yield rather ill-behaved statistical models, there are subclasses of models with correlated errors in which some of the nice properties of DAG models are preserved; compare McDonald (2002). In this paper we consider path diagrams in which there are no directed cycles and no 'double' edges of the form $i \leftrightarrows j$ (compare Def. 2 and 3). Since such double edges have been called 'bows', we call this class *bow-free acyclic path diagrams* (BAPs). An example of a BAP arose in our motivating example in §1.1; see Figure 1. While instrumental variable models, which are much studied in economics, contain bows, most models in other social sciences are based on BAPs. For instance, all path diagrams in Bollen (1989) are BAPs.

Bow-free acyclic path diagrams were also considered by Brito and Pearl (2002) who showed that the associated normal linear models are almost everywhere identifiable; see §2.2 for the definition. This result and other identification properties of BAP models are reviewed in Section 2. In Section 3 we give details on likelihood equations and Fisher-information of normal structural equation models. This sets the scene for our main contribution: the *Residual Iterative Conditional Fitting* (RICF) algorithm for maximization of the likelihood function of BAP models, which is presented in Section 4. Standard software for structural equation modelling currently employs general-purpose optimization routines for this task (Bollen, 1989, Appendix 4C). Many of these algorithms, however, neglect constraints of positive definiteness on the covariance matrix and suffer from convergence problems. According to Steiger (2001), failure to converge is 'not uncommon' and presents significant challenges to novice users of existing software. In contrast, our RICF algorithm produces positive definite covariance matrix estimates during all its iterations and has good convergence properties, as illustrated in the simulations in Section 5. Further discussion of RICF is provided in Section 6.

2. Normal Linear Models and Path Diagrams

Let $Y = (Y_i \mid i \in V) \in \mathbb{R}^V$ be a random vector, indexed by the finite set V, that follows a multivariate normal distribution $\mathcal{N}(0,\Sigma)$ with positive definite covariance matrix Σ. A zero mean vector is assumed merely to avoid notational overhead. The models we consider subsequently are induced by linear structural equations as follows.

2.1 Systems of Structural Equations and Path Diagrams

Let $\{\mathrm{pa}(i) \mid i \in V\}$ and $\{\mathrm{sp}(i) \mid i \in V\}$ be two families of index sets. For reasons explained below, we refer to these index sets as sets of parents and spouses, respectively. We require that $i \notin \mathrm{pa}(i) \cup \mathrm{sp}(i)$ for all $i \in V$; moreover, let the second family satisfy the symmetry condition that $j \in \mathrm{sp}(i)$ if and only if $i \in \mathrm{sp}(j)$. These two families determine a system of structural equations

$$Y_i = \sum_{j \in \mathrm{pa}(i)} \beta_{ij} Y_j + \varepsilon_i, \quad i \in V, \tag{5}$$

whose zero-mean errors ε_i and ε_j are uncorrelated if $i \notin \mathrm{sp}(j)$, or equivalently, $j \notin \mathrm{sp}(i)$. The equations in (5) correspond to a *path diagram*, that is, a mixed graph G featuring both *directed* ($\rightarrow$) and *bi-directed* ($\leftrightarrow$) edges but no edges from a vertex i to itself (see Figures 1 and 2). The vertex set of G is the index set V, and G contains the edge $j \rightarrow i$ if and only if $j \in \mathrm{pa}(i)$, and the edge $j \leftrightarrow i$ if and only if $j \in \mathrm{sp}(i)$ (or equivalently, $i \in \mathrm{sp}(j)$). Subsequently, we exploit the path diagram representation of (5). If $i \rightarrow j$ is an edge in G, then we call i a *parent* of j, and if $i \leftrightarrow j$ is in G then i is referred to as a *spouse* of j. Thus, as remarked above, $\mathrm{pa}(i)$, $\mathrm{sp}(i)$ are, respectively, the sets of parents and spouses of i.

Let G be a path diagram and define $\mathbf{B}(G)$ to be the collection of all $V \times V$ matrices $B = (\beta_{ij})$ that satisfy

$$\beta_{ij} = 0 \quad \text{whenever } j \rightarrow i \text{ is not an edge in } G, \tag{6}$$

and are such that $I - B$ is invertible. Let $\mathbf{P}(V)$ be the cone of positive definite $V \times V$ matrices and $\mathbf{O}(G) \subseteq \mathbf{P}(V)$ the set of matrices $\Omega = (\omega_{ij}) \in \mathbf{P}(V)$ that satisfy

$$\omega_{ij} = 0 \quad \text{whenever } i \neq j \text{ and } j \leftrightarrow i \text{ is not in } G. \tag{7}$$

(Here and in the sequel, a symbol such as V denotes both a finite set and its cardinality.) The system (5) associated with the path diagram G can be written compactly as $Y = BY + \varepsilon$. If we assume that $B \in \mathbf{B}(G)$ and that the error covariance matrix $\mathrm{Var}(\varepsilon) = \Omega$ is in $\mathbf{O}(G)$, then (5) has a unique solution Y that is a multivariate normal random vector with covariance matrix $\Sigma = (I-B)^{-1}\Omega(I-B)^{-t}$. Here, I is the identity matrix and the superscript '$-t$' stands for transposition and inversion.

The above considerations lead to the following definition of a linear model associated with a path diagram (or equivalently, a system of structural equations).

Definition 1 *The normal linear model $\mathbf{N}(G)$ associated with a path diagram G is the family of multivariate normal distributions $\mathcal{N}(0,\Sigma)$ with covariance matrix in the set $\mathbf{P}(G) = \left\{(I-B)^{-1}\Omega(I-B)^{-t} \mid B \in \mathbf{B}(G), \Omega \in \mathbf{O}(G)\right\}$. We call the map $\Phi_G : \mathbf{B}(G) \times \mathbf{O}(G) \rightarrow \mathbf{P}(G)$ given by*

$$\Phi_G(B,\Omega) = (I-B)^{-1}\Omega(I-B)^{-t}$$

the parameterization map of $\mathbf{N}(G)$.

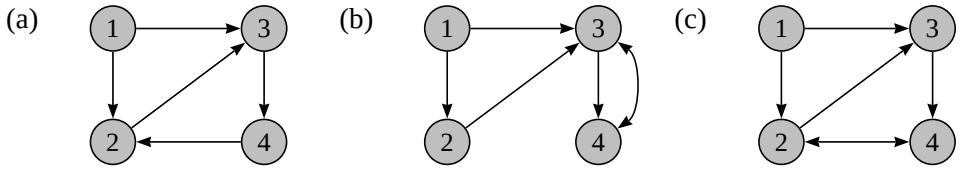

Figure 2: Path diagrams that are (a) cyclic, (b) acyclic but not bow-free, (c) acyclic and bow-free. Only path diagram (c) yields a curved exponential family.

Example 1 The path diagram G in Figure 2(a) depicts the equation system

$$\begin{aligned} Y_1 &= \varepsilon_1, & Y_2 &= \beta_{21}Y_1 + \beta_{24}Y_4 + \varepsilon_2, \\ Y_3 &= \beta_{31}Y_1 + \beta_{32}Y_2 + \varepsilon_3, & Y_4 &= \beta_{43}Y_3 + \varepsilon_4, \end{aligned}$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ are pairwise uncorrelated, that is, the matrices $\Omega \in \mathbf{O}(G)$ are diagonal. This system exhibits a circular covariate-response structure as the path diagram contains the directed cycle $2 \to 3 \to 4 \to 2$. This feedback loop is reflected in the fact that $\det(I - B) = 1 - \beta_{24}\beta_{43}\beta_{32}$ for $B \in \mathbf{B}(G)$. Therefore, the path coefficients need to satisfy $\beta_{24}\beta_{43}\beta_{32} \neq 1$ in order to lead to a positive definite covariance matrix in $\mathbf{P}(G)$. This example is considered in more detail in Drton (2009), where it is shown that the parameter space $\mathbf{P}(G)$ has singularities that lead to non-standard behavior of likelihood ratio tests.

The models considered in the remainder of this paper do not have any feedback loops, that is, they have the following structure.

Definition 2 *A path diagram G and its associated normal linear model $\mathbf{N}(G)$ are* recursive *or* acyclic *if G does not contain directed cycles, that is, there do not exist $i, i_1, \ldots, i_k \in V$ such that G features the edges $i \to i_1 \to i_2 \to \cdots \to i_k \to i$.*

We use the term *acyclic* rather than *recursive*, as some authors have used the term 'recursive' for path diagrams that are acyclic *and* contain no bi-directed edges. If G is acyclic, then the vertices in V can be ordered such that a matrix B that satisfies (6) is lower-triangular. It follows that

$$\det(I - B) = 1. \tag{8}$$

In particular, $I - B$ is invertible for any choice of the path coefficients β_{ij}, $j \to i$ in G, and the parameterization map Φ_G is a polynomial map.

2.2 Bow-free Acyclic Path Diagrams (BAPs)

The normal linear model $\mathbf{N}(G)$ associated with a path diagram G is *everywhere identifiable* if the parameterization map Φ_G is one-to-one, that is, for all $B_0 \in \mathbf{B}(G)$ and $\Omega_0 \in \mathbf{O}(G)$ it holds that

$$\Phi_G(B, \Omega) = \Phi_G(B_0, \Omega_0) \quad \Longrightarrow \quad B = B_0 \text{ and } \Omega = \Omega_0. \tag{9}$$

If there exists a Lebesgue null set $N_G \subseteq \mathbf{B}(G) \times \mathbf{O}(G)$ such that (9) holds for all $(B_0, \Omega_0) \notin N_G$, then we say that $\mathbf{N}(G)$ is *almost everywhere identifiable*.

Acyclic path diagrams may contain *bows*, that is, double edges $i \leftrightarrows j$. It is easy to see that normal linear models associated with path diagrams with bows are never everywhere identifiable. However, they may sometimes be almost everywhere identifiable as is the case for the next example. This example illustrates that almost everywhere identifiability is not enough to ensure regular behavior of statistical procedures.

Example 2 The path diagram G in Figure 2(b) features the bow $3 \leftrightarrows 4$. The associated normal linear model $\mathbf{N}(G)$ is also known as an instrumental variable model. The 9-dimensional parameter space $\mathbf{P}(G)$ is part of the hypersurface defined by the vanishing of the so-called *tetrad* $\sigma_{13}\sigma_{24} - \sigma_{14}\sigma_{23}$. It follows that the model $\mathbf{N}(G)$ lacks regularity because the tetrad hypersurface has singularities at points $\Sigma \in \mathbf{P}(G)$ with $\sigma_{13} = \sigma_{14} = \sigma_{23} = \sigma_{24} = 0$. These singularities occur if and only if $\beta_{31} = \beta_{32} = 0$, and correspond to points at which the identifiability property in (9) fails to hold. This lack of smoothness expresses itself statistically, for example, when testing the hypothesis $\beta_{31} = \beta_{32} = 0$ in model $\mathbf{N}(G)$. Using the techniques in Drton (2009), the likelihood ratio statistic for this problem can be shown to have non-standard behavior with a large-sample limiting distribution that is given by the larger of the two eigenvalues of a 2×2-Wishart matrix with 2 degrees of freedom and the identity matrix as scale parameter.

Definition 3 *A path diagram G and its associated normal linear model $\mathbf{N}(G)$ are* bow-free *if G contains at most one edge between any pair of vertices. If G is bow-free and acyclic, we call it a bow-free acyclic path diagram (BAP).*

As stressed in the introduction, BAPs are widespread in applications. Examples are shown in Figures 1, 2(c) and 6. Contrary to some path diagrams with bows, the normal linear models associated with BAPs are always at least almost everywhere identifiable.

Theorem 4 (Brito and Pearl, 2002) *If G is a BAP, then the normal linear model $\mathbf{N}(G)$ is almost everywhere identifiable.*

Many BAP models are in fact everywhere identifiable.

Theorem 5 (Richardson and Spirtes, 2002) *Suppose G is an ancestral BAP, that is, G does not contain an edge $i \leftrightarrow j$ such that there is a directed path $j \to i_1 \to \cdots \to i_k \to i$ that leads from vertex j to vertex i. Then the normal linear model $\mathbf{N}(G)$ is everywhere identifiable.*

The next example shows that the condition in Theorem 5 is sufficient but not necessary for identification. The characterization of the class of BAPs whose associated normal linear models are everywhere identifiable remains an open problem.

Example 3 The BAP G in Figure 2(c) is not ancestral because it contains the edges $4 \leftrightarrow 2 \to 3 \to 4$. Nevertheless, the associated normal linear model $\mathbf{N}(G)$ is everywhere identifiable, which can be shown by identifying the parameters in B and Ω row-by-row following the order $1 < 2 < 3 < 4$. It is noteworthy that the model $\mathbf{N}(G)$ in this example is not a Markov model, that is, a generic multivariate normal distribution in $\mathbf{N}(G)$ exhibits no conditional independence relations. Instead, the entries of covariance matrices $\Sigma = (\sigma_{ij}) \in \mathbf{P}(G)$ satisfy

$$(\sigma_{11}\sigma_{22} - \sigma_{12}^2)(\sigma_{14}\sigma_{33} - \sigma_{13}\sigma_{34}) = (\sigma_{13}\sigma_{24} - \sigma_{14}\sigma_{23})(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23}). \tag{10}$$

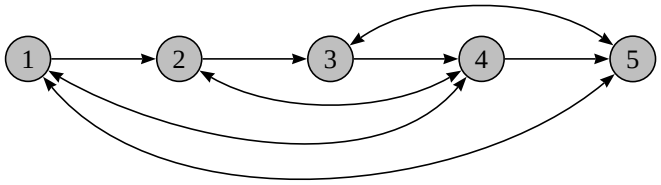

Figure 3: Bow-free acyclic path diagram whose associated normal linear model is almost, but not everywhere, identifiable. The model is not a curved exponential family.

The constraint in (10) has a nice interpretation. Let $(Y_1,\ldots,Y_4)$ have (positive definite) covariance matrix $\Sigma = (\sigma_{ij})$, and define $e_2 = Y_2 - \sigma_{21}/\sigma_{11}Y_1$ to be the residual in the regression of Y_2 on Y_1. Then (10) holds for Σ if and only if Y_1 and Y_4 are conditionally independent given e_2 and Y_3.

The above-stated Theorem 4 was proved in Brito and Pearl (2002), and an inspection of their proof reveals the following fact.

Lemma 6 *If the normal linear model* $\mathbf{N}(G)$ *associated with a BAP G is everywhere identifiable, then the (bijective) parameterization map* Φ_G *has an inverse that is a rational map with no pole on* $\mathbf{P}(G)$.

By (8), the parameterization map Φ_G for a BAP G is polynomial and thus smooth. If Φ_G^{-1} is rational and without pole, then the image of Φ_G, that is, $\mathbf{P}(G)$ is a smooth manifold (see, e.g., Edwards, 1994, II.4). This has an important consequence.

Corollary 7 *If the normal linear model* $\mathbf{N}(G)$ *associated with a BAP G is everywhere identifiable, then* $\mathbf{N}(G)$ *is a curved exponential family.*

The theory of curved exponential families is discussed by Kass and Vos (1997). It implies in particular that maximum likelihood estimators in curved exponential families are asymptotically normal, and that likelihood ratio statistics comparing two such families are asymptotically chi-square regardless of where in the null hypothesis a true parameter is located. Unfortunately, however, Lemma 6 and Corollary 7 do not hold for every BAP.

Example 4 The normal linear model associated with the BAP G in Figure 3 is not a curved exponential family. In this model the identifiability property in (9) breaks down if and only if (B,Ω) satisfy

$$\beta_{21}\omega_{14}\omega_{24} - \omega_2\omega_4 + \omega_{24}^2 = 0, \quad \beta_{32}\beta_{43}\omega_2 + \omega_{24} = 0.$$

It can be shown that the covariance matrices $\Phi_G(B,\Omega)$ associated with this set of parameters yield points at which the 13-dimensional set $\mathbf{P}(G)$ has more than 13 linearly independent tangent directions. Hence, $\mathbf{P}(G)$ is singular at these covariance matrices.

3. Likelihood Inference

Suppose a sample of size N is drawn from a multivariate normal distribution $\mathcal{N}(0,\Sigma)$ in the linear model $\mathbf{N}(G)$ associated with a BAP $G=(V,E)$. We group the observed random vectors as columns in the $V\times N$ matrix Y such that Y_{in} represents the observation of variable i on subject n. Having assumed a zero mean vector, we define the empirical covariance matrix as

$$S=(s_{ij})=\frac{1}{N}YY^t\in\mathbb{R}^{V\times V}.$$

Assuming $N\geq V$, the matrix S is positive definite with probability one. (As before, V here also denotes the cardinality of the set.) Models with unknown mean vector $\mu\in\mathbb{R}^V$ can be treated by estimating μ by the empirical mean vector and adjusting the empirical covariance matrix accordingly; $N\geq V+1$ then ensures almost sure positive definiteness of the empirical covariance matrix.

3.1 Likelihood Function and Likelihood Equations

Given observations Y with empirical covariance matrix S, the log-likelihood function $\ell:\mathbf{B}(G)\times\mathbf{O}(G)\to\mathbb{R}$ of the model $\mathbf{N}(G)$ takes the form

$$\ell(B,\Omega)=-\frac{N}{2}\log\det(\Omega)-\frac{N}{2}\operatorname{tr}\big[(I-B)^t\Omega^{-1}(I-B)S\big]. \tag{11}$$

Here we ignored an additive constant and used that $\det(I-B)=1$ if $B\in\mathbf{B}(G)$; compare (8). Let $\beta=(\beta_{ij}\mid i\in V,\ j\in\mathrm{pa}(i))$ and $\omega=(\omega_{ij}\mid i\leq j,\ j\in\mathrm{sp}(i)\text{ or }i=j)$ be the vectors of unconstrained elements in B and Ω. Let P and Q be the matrices with entries in $\{0,1\}$ that satisfy $\mathrm{vec}(B)=P\beta$ and $\mathrm{vec}(\Omega)=Q\omega$, respectively, where $\mathrm{vec}(A)$ refers to stacking of the columns of the matrix A. Taking the first derivatives of $\ell(B,\Omega)$ with respect to β and ω we obtain the likelihood equations.

Proposition 8 *The likelihood equations of the normal linear model $\mathbf{N}(G)$ associated with a BAP G can be written as*

$$P^t\operatorname{vec}\big(\Omega^{-1}(I-B)S\big)=P^t\operatorname{vec}\big(\Omega^{-1}S\big)-P^t\big(S\otimes\Omega^{-1}\big)P\beta=0, \tag{12}$$

$$Q^t\operatorname{vec}\big(\Omega^{-1}-\Omega^{-1}(I-B)S(I-B)^t\Omega^{-1}\big)=0, \tag{13}$$

where $\otimes$ denotes the Kronecker product.

In general, the likelihood equations need to be solved iteratively. One possible approach proceeds by alternately solving (12) and (13) for β and ω, respectively. For fixed ω, (12) is a linear equation in β and easily solved. For fixed β, (13) constitutes the likelihood equations of a multivariate normal covariance model for $\varepsilon=(I-B)Y$, which is specified by requiring that $\Omega_{ij}=0$ whenever the edge $i\leftrightarrow j$ is not in G. The solution of (13), with β fixed, requires, in general, another iterative method. As an alternative to this nesting of two iterative methods, we propose in Section 4 a method that solves (12) and (13) in joint updates of β and ω.

Remark 9 When proving their identifiability result for BAP models, Brito and Pearl (2002) gave an algorithm for recovering the parameters β and ω from a population covariance matrix. Applied to the empirical covariance matrix S, this algorithm produces consistent estimates $\tilde\beta$ and $\tilde\omega$. However, these are generally not the maximum likelihood estimators (MLE) and the error covariance matrix corresponding to $\tilde\omega$ may fail to be positive definite.

3.2 Fisher-Information

Large-sample confidence intervals for (β, ω) can be obtained by approximating the distribution of the MLE $(\hat{\beta}, \hat{\omega})$ by the normal distribution with mean vector (β, ω) and covariance matrix $\frac{1}{N} I(\beta, \omega)^{-1}$. Here, $I(\beta, \omega)$ denotes the Fisher-information, which, as shown in Appendix A, is of the following form.

Proposition 10 *The (expected) Fisher-information of the normal linear model $\mathbf{N}(G)$ associated with a BAP G is*

$$I(\beta, \omega) = \begin{pmatrix} P^t\big(\Sigma \otimes \Omega^{-1}\big)P & P^t\big[(I-B)^{-1} \otimes \Omega^{-1}\big]Q \\ Q^t\big[(I-B)^{-t} \otimes \Omega^{-1}\big]P & \frac{1}{2} Q^t\big(\Omega^{-1} \otimes \Omega^{-1}\big)Q \end{pmatrix}.$$

The Fisher-information in Proposition 10 need not be block-diagonal, in which case the estimation of the covariances ω affects the asymptotic variance of the MLE $\hat{\beta}$. However, this does not happen for *bi-directed chain graphs*, which form one of the model classes discussed by Wermuth and Cox (2004). A path diagram G is a bi-directed chain graph if its vertex set V can be partitioned into disjoint subsets $\tau_1, \ldots, \tau_T$, known as *chain components*, such that all edges in each subgraph G_{τ_t} are bi-directed and edges between two subsets $\tau_s \neq \tau_t$ are directed, pointing from τ_s to τ_t, if $s < t$. Since bi-directed chain graphs are ancestral graphs the associated normal linear models are everywhere identifiable.

Proposition 11 *For a BAP G, the following two statements are equivalent:*

(i) For all underlying covariance matrices $\Sigma \in \mathbf{P}(G)$, the MLEs of the parameter vectors β and ω of the normal linear model $\mathbf{N}(G)$ are asymptotically independent.

(ii) The path diagram G is a bi-directed chain graph.

A proof of Proposition 11 is given in Appendix A. This result is an instance of the asymptotic independence of mean and natural parameters in mixed parameterizations of exponential families (Barndorff-Nielsen, 1978).

4. Residual Iterative Conditional Fitting

We now present an algorithm for computing the MLE in the normal linear model $\mathbf{N}(G)$ associated with a BAP. The algorithm extends the *iterative conditional fitting* (ICF) procedure of Chaudhuri et al. (2007), which is for path diagrams with exclusively bi-directed edges.

Let $Y_i \in \mathbb{R}^N$ denote the i-th row of the observation matrix Y and $Y_{-i} = Y_{V \setminus \{i\}}$ the $(V \setminus \{i\}) \times N$ submatrix of Y. The ICF algorithm proceeds by repeatedly iterating through all vertices $i \in V$ and carrying out three steps: (i) fix the marginal distribution of Y_{-i}, (ii) fit the conditional distribution of Y_i given Y_{-i} under the constraints implied by the model $\mathbf{N}(G)$, and (iii) obtain a new estimate of Σ by combining the estimated conditional distribution $(Y_i \mid Y_{-i})$ with the fixed marginal distribution of Y_{-i}. The crucial point is then that for path diagrams containing only bi-directed edges, the problem of fitting the conditional distribution for $(Y_i \mid Y_{-i})$ under the constraints of the model can be rephrased as a least squares regression problem. Unfortunately, the consideration of the conditional distribution of $(Y_i \mid Y_{-i})$ is complicated for path diagrams that contain also directed edges. However, as we show below, the directed edges can be 'removed' by consideration of *residuals*, which here refers to estimates of the error terms $\varepsilon = (I - B)Y$. Since it is based on this idea, we give our new extended algorithm the name *Residual Iterative Conditional Fitting* (RICF).

4.1 The RICF Algorithm

The main building block of the new algorithm is the following decomposition of the log-likelihood function. We adopt the shorthand notation X_C for the $C \times N$ submatrix of a $D \times N$ matrix X, where $C \subseteq D$.

Theorem 12 *Let G be a BAP and $i \in V$. Let $\|x\|^2 = x^t x$ and define*

$$\omega_{ii.-i} = \omega_{ii} - \Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i} \tag{14}$$

to be the conditional variance of ε_i given ε_{-i}; recall that $\Omega^{-1}_{-i,-i} = (\Omega_{-i,-i})^{-1}$. Then the log-likelihood function $\ell(B,\Omega)$ of the model $\mathbf{N}(G)$ can be decomposed as

$$\begin{aligned}\ell(B,\Omega) = &-\frac{N}{2}\log\omega_{ii.-i} - \frac{1}{2\omega_{ii.-i}}\left\|Y_i - B_{i,\mathrm{pa}(i)}Y_{\mathrm{pa}(i)} - \Omega_{i,\mathrm{sp}(i)}(\Omega^{-1}_{-i,-i}\varepsilon_{-i})_{\mathrm{sp}(i)}\right\|^2 \\ &-\frac{N}{2}\log\det(\Omega_{-i,-i}) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}_{-i,-i}\varepsilon_{-i}\varepsilon^t_{-i}\right).\end{aligned} \tag{15}$$

Proof Forming $\varepsilon = (I-B)Y$, we rewrite (11) as

$$\ell(B,\Omega) = -\frac{N}{2}\log\det(\Omega) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}\varepsilon\varepsilon^t\right) =: \ell(\Omega \mid \varepsilon). \tag{16}$$

Using the inverse variance lemma (Whittaker, 1990, Prop. 5.7.3), we partition Ω^{-1} as

$$\begin{pmatrix}\omega_{ii} & \Omega_{i,-i}\\ \Omega_{-i,i} & \Omega_{-i,-i}\end{pmatrix}^{-1} = \begin{pmatrix}\omega^{-1}_{ii.-i} & -\omega^{-1}_{ii.-i}\Omega_{i,-i}\Omega^{-1}_{-i,-i}\\ -\Omega^{-1}_{-i,-i}\Omega_{-i,i}\omega^{-1}_{ii.-i} & \Omega^{-1}_{-i,-i} + \Omega^{-1}_{-i,-i}\Omega_{-i,i}\omega^{-1}_{ii.-i}\Omega_{i,-i}\Omega^{-1}_{-i,-i}\end{pmatrix}.$$

We obtain that the log-likelihood function in (16) equals

$$\begin{aligned}\ell(\Omega \mid \varepsilon) = &-\frac{N}{2}\log\omega_{ii.-i} - \frac{1}{2\omega_{ii.-i}}\left\|\varepsilon_i - \Omega_{i,-i}\Omega^{-1}_{-i,-i}\varepsilon_{-i}\right\|^2 \\ &-\frac{N}{2}\log\det(\Omega_{-i,-i}) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}_{-i,-i}\varepsilon_{-i}\varepsilon^t_{-i}\right).\end{aligned}$$

By definition, $\varepsilon_i = Y_i - B_{i,\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}$. Moreover, under the restrictions (7),

$$\Omega_{i,-i}\Omega^{-1}_{-i,-i}\varepsilon_{-i} = \Omega_{i,\mathrm{sp}(i)}(\Omega^{-1}_{-i,-i}\varepsilon_{-i})_{\mathrm{sp}(i)},$$

which yields the claimed decomposition. ■

The log-likelihood decomposition (15) is essentially based on the decomposition of the joint distribution of ε into the marginal distribution of ε_{-i} and the conditional distribution $(\varepsilon_i \mid \varepsilon_{-i})$. This leads to the idea of building an iterative algorithm whose steps are based on fixing the marginal distribution of ε_{-i} and estimating a conditional distribution. In order to fix the marginal distribution ε_{-i} we need to fix the submatrix $\Omega_{-i,-i}$ comprising all but the i-th row and column of Ω as well as the submatrix $B_{-i,V}$, which comprises all but the i-th row of B. With $\Omega_{-i,-i}$ and $B_{-i,V}$ fixed we can compute ε_{-i} as well as the *pseudo-variables*, defined by

$$Z_{-i} = \Omega^{-1}_{-i,-i}\varepsilon_{-i}. \tag{17}$$

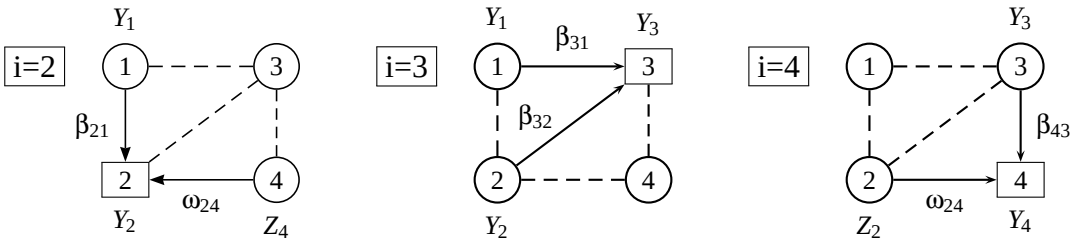

Figure 4: Illustration of the RICF update steps in Example 5. The structure of each least squares regression is indicated by directed edges pointing from the predictor variables to the response variable depicted by a square node. (See text for details.)

From (15), it now becomes apparent that, for fixed $\Omega_{-i,-i}$ and $B_{-i,V}$, the maximization of the log-likelihood function $\ell(B,\Omega)$ can be solved by maximizing the function

$$\big((\beta_{ij})_{j\in \mathrm{pa}(i)}, (\omega_{ik})_{k\in \mathrm{sp}(i)}, \omega_{ii.-i}\big) \mapsto - \frac{N}{2}\log \omega_{ii.-i} - \frac{1}{2\omega_{ii.-i}}\Big\|Y_i - \sum_{j\in \mathrm{pa}(i)} \beta_{ij} Y_j - \sum_{k\in \mathrm{sp}(i)} \omega_{ik} Z_k\Big\|^2 \quad (18)$$

over $\mathbb{R}^{\mathrm{pa}(i)} \times \mathbb{R}^{\mathrm{sp}(i)} \times (0,\infty)$. The maximizers of (18), however, are the least squares estimates in the regression of Y_i on both $(Y_j \mid j \in \mathrm{pa}(i))$ and $(Z_k \mid k \in \mathrm{sp}(i))$.

Employing the above observations, the *RICF algorithm* for computing the MLE $(\hat{B},\hat{\Omega})$ repeats the following steps for each $i \in V$:

1. Fix $\Omega_{-i,-i}$ and $B_{-i,V}$, and compute residuals ε_{-i} and pseudo-variables $Z_{\mathrm{sp}(i)}$;
2. Obtain least squares estimates of β_{ij}, $j \in \mathrm{pa}(i)$, ω_{ik}, $k \in \mathrm{sp}(i)$, and $\omega_{ii.-i}$ by regressing response variable Y_i on the covariates Y_j, $j \in \mathrm{pa}(i)$ and Z_k, $k \in \mathrm{sp}(i)$;
3. Compute an estimate of $\omega_{ii} = \omega_{ii.-i} + \Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i}$ using the new estimates and the fixed parameters; compare (14).

After steps 1 to 3, we move on to the next vertex in V. After the last vertex in V we return to consider the first vertex. The procedure is continued until convergence.

Example 5 For illustration of the regressions performed in RICF, we consider the normal linear model associated with the BAP G in Figure 2(c). The parameters to be estimated in this model are β_{21}, β_{31}, β_{32}, β_{43} and ω_{11}, ω_{22}, ω_{33}, ω_{44}, ω_{24}.

Vertex 1 in Figure 2(c) has no parents or spouses, and its RICF update step consists of a trivial regression. In other words, the variance ω_{11} is the unconditional variance of Y_1 with MLE $\hat{\omega}_{11} = s_{11}$. For the remaining vertices, the corresponding RICF update steps are illustrated in Figure 4, where the response variable Y_i in the i-th update step is shown as a square node while the remaining variables are depicted as circles. Directed edges indicate variables acting as covariates in the least squares regression. These covariates are labelled according to whether the random variable Y_j, or the pseudo-variable Z_j defined in (17), is used in the regression. Note that since $\mathrm{sp}(3) = \emptyset$, repetition of steps 1-3 in §4.1 is required only for $i \in \{2,4\}$.

In RICF, the log-likelihood function $\ell(B,\Omega)$ from (11) is repeatedly maximized over sections in the parameter space defined by fixing the parameters $\Omega_{-i,-i}$, and $B_{-i,V}$. RICF thus is an iterative partial maximization algorithm and has the following convergence properties.

Theorem 13 *If G is a BAP and the empirical covariance matrix S is positive definite, then the following holds:*

(i) For any starting value $(\hat{B}^0,\hat{\Omega}^0) \in \mathbf{B}(G) \times \mathbf{O}(G)$, RICF constructs a sequence of estimates $(\hat{B}^s,\hat{\Omega}^s)_s$ in $\mathbf{B}(G) \times \mathbf{O}(G)$ whose accumulation points are local maxima or saddle points of the log-likelihood function $\ell(B,\Omega)$. Moreover, evaluating the log-likelihood function at different accumulation points yields the same value.

(ii) If the normal linear model $\mathbf{N}(G)$ is everywhere identifiable and the likelihood equations have only finitely many solutions then the sequence $(\hat{B}^s,\hat{\Omega}^s)_s$ converges to one of these solutions.

Proof Let $\ell(\Sigma)$ be the log-likelihood function for the model of all centered multivariate normal distributions on $\mathbb{R}^V$. If S is positive definite then the set C that comprises all positive definite matrices $\Sigma \in \mathbb{R}^{V\times V}$ at which $\ell(\Sigma) \geq \ell(\hat{B}^0,\hat{\Omega}^0)$ is compact. In particular, the log-likelihood function in (11) is bounded, and claim (i) can be derived from general results about iterative partial maximization algorithms; see for example, Drton and Eichler (2006). For claim (ii) note that if $\mathbf{N}(G)$ is everywhere identifiable, then the compact set C has compact preimage $\phi_G^{-1}(C)$ under the model parameterization map; recall Lemma 6. ■

Remark 14 If the normal linear model $\mathbf{N}(G)$ associated with a BAP G is not everywhere identifiable, then it is possible that a sequence of estimates $(\hat{B}^s,\hat{\Omega}^s)_s$ produced by RICF diverges and does not have any accumulation points. In these cases, however, the corresponding sequence of covariance matrices $\Phi_G(\hat{B}^s,\hat{\Omega}^s)_s$ still has at least one accumulation point because it ranges in the compact set C exhibited in the proof of Theorem 13. Divergence of $(\hat{B}^s,\hat{\Omega}^s)_s$ occurs in two instances in the simulations in §5; compare Table 1. In both cases, the sequence $\Phi_G(\hat{B}^s,\hat{\Omega}^s)_s$ converges to a positive definite covariance matrix.

4.2 Computational Savings in RICF

If G is a DAG, that is, an acyclic path diagram without bi-directed edges, then the MLE $(\hat{B},\hat{\Omega})$ in $\mathbf{N}(G)$ can be found in a finite number of regressions (e.g., Wermuth, 1980). However, we can also run RICF. Since in a DAG, $\mathrm{sp}(i) = \emptyset$ for all $i \in V$, step 2 of RICF regresses variable Y_i solely on its parents Y_j, $j \in \mathrm{pa}(i)$. Not involving pseudo-variables that could change from one iteration to the other, this regression remains the same throughout different iterations, and RICF converges in one step.

Similarly, for a general BAP G, if vertex $i \in V$ has no spouses, $\mathrm{sp}(i) = \emptyset$, then the MLE of $B_{i,\mathrm{pa}(i)}$ and ω_{ii} can be determined by a single iteration of the algorithm. In other words, RICF reveals these parameters as being estimable in closed form, namely as rational functions of the data. (This occurred for vertex $i = 3$ in Example 5.) It follows that, to estimate the remaining parameters, the iterations need only be continued over vertices j with $\mathrm{sp}(j) \neq \emptyset$.

For further computational savings note that $\Omega_{\mathrm{dis}(i),V\setminus(\mathrm{dis}(i)\cup\{i\})} = 0$, where $\mathrm{dis}(i) = \{j \mid j \leftrightarrow \cdots \leftrightarrow i, j \neq i\}$ is the district of $i \in V$. Hence, since $\mathrm{sp}(i) \subseteq \mathrm{dis}(i)$,

$$(\Omega^{-1}_{-i,-i}\varepsilon_{-i})_{\mathrm{sp}(i)} = (\Omega^{-1}_{\mathrm{dis}(i),\mathrm{dis}(i)}\varepsilon_{\mathrm{dis}(i)})_{\mathrm{sp}(i)};$$

see Koster (1999, Lemma 3.1.6) and Richardson and Spirtes (2002, Lemma 8.10). Since $\varepsilon_{\mathrm{dis}(i)} = Y_{\mathrm{dis}(i)} - B_{\mathrm{dis}(i),\mathrm{pa}(\mathrm{dis}(i))}Y_{\mathrm{pa}(\mathrm{dis}(i))}$, it follows that in the RICF update step for vertex i attention can be restricted to the variables in $\{i\} \cup \mathrm{pa}(i) \cup \mathrm{dis}(i) \cup \mathrm{pa}(\mathrm{dis}(i))$.

Finally, note that while the RICF algorithm is described in terms of the entire data matrix Y, the least squares estimates computed in its iterations are clearly functions of the empirical covariance matrix, which is a sufficient statistic.

5. Simulation Studies

In order to evaluate the performance of the RICF algorithm we consider two scenarios. First, we fit linear models based on randomly generated BAPs to gene expression data. This scenario is relevant for model selection tasks, and we compare RICF's performance in this problem to that of algorithms implemented in software for structural equation modelling. Second, we study how RICF behaves when it is used to fit larger models to data simulated from the respective model. In contrast to the first scenario, the second scenario involves models that generally fit the considered data well.

5.1 Gene Expression Data

We consider data on gene expression in *Arabidopsis thaliana* from Wille et al. (2004). We restrict attention to 13 genes that belong to one pathway: DXPS1-3, DXR, MCT, CMK, MECPS, HDS, HDR, IPPI1, GPPS, PPDS1-2. Data from $n = 118$ microarray experiments are available. We fit randomly generated BAP models to these data using RICF and two alternative methods.

The BAP models are generated as follows. For each of the 78 possible pairs of vertices $i < j$ in $V = \{1,\ldots,13\}$ we draw from a multinomial distribution to generate a possible edge. The probability for drawing the edge $i \to j$ is d, and the probability for drawing $i \leftrightarrow j$ is b so that with probability $1-d-b$ there is no edge between i and j. We then apply a random permutation to the vertices to obtain the final BAP. For each of twelve combinations (d,b) with $d = 0.05, 0.1, 0.2, 0.3$ and $b = 0.05, 0.1, 0.2$, we simulate 1000 BAPs. The expected number of edges thus varies between 7.8 and 39.

For fitting the simulated BAPs to the gene expression data, we implemented RICF in the statistical programming environment R (R Development Core Team, 2007). As alternatives, we consider the R package 'sem' (Fox, 2006) and the widely used software LISREL (Jöreskog and Sörbom, 1997) in its student version 8.7 for Linux (student versions are free but limited to 15 variables). Both these programs employ general purpose optimizers, for example, 'sem' makes a call to the R function 'nlm'.

Our simulation results are summarized in Table 1. Each row in the table corresponds to a choice of the edge probabilities d and b. The first three columns count how often, in 1000 simulations, the three considered fitting routines failed to converge. The starting values of LISREL and 'sem' were set according to program defaults, and RICF was started by setting $\hat{B}^{(0)}$ and $\hat{\Omega}^{(0)}$ equal to the MLE in the DAG model associated with the DAG obtained by removing all bi-directed edges from the considered BAP.

		No convergence			All	All	Running time		
d	*b*	RICF	LIS	SEM	converge	agree	RICF	LIS	SEM
0.05	0.05	0	36	47	941	940	0.03	0.02	1.15
	0.1	0	177	221	746	739	0.09	0.03	1.58
	0.2	0	499	599	347	333	0.21	0.04	2.71
0.1	0.05	0	32	36	951	949	0.04	0.03	1.58
	0.1	0	137	193	786	780	0.09	0.03	2.09
	0.2	0	440	610	364	354	0.25	0.04	3.43
0.2	0.05	0	19	39	958	954	0.05	0.03	2.67
	0.1	0	91	176	815	808	0.13	0.04	3.34
	0.2	1	326	520	461	452	0.33	0.05	5.03
0.3	0.05	0	16	38	960	957	0.06	0.04	4.04
	0.1	0	59	136	859	850	0.17	0.04	4.96
	0.2	1	225	471	519	490	0.40	0.06	6.97

Table 1: Fitting simulated BAPs to gene expression data using RICF, LISREL and 'sem'. Each row is based on 1000 simulations. Running time is average CPU time (in sec.) for the cases in which all three algorithms converged. (See text for details.)

LISREL and 'sem' failed to converge for a rather large number of models. The LISREL output explained why convergence failed, and virtually all failures were due to the optimizer converging to matrices that were not positive definite. The remedy would be to try new starting values but doing this successfully in an automated fashion is a challenging problem in itself. For RICF convergence failure arose in only two cases. In both cases the RICF estimates $(\hat{B},\hat{\Omega})$ had some diverging entries. Despite the divergence in (B,Ω)-space, the sequence of associated covariance matrices $\Phi_G(\hat{B},\hat{\Omega})$ computed by RICF converged, albeit very slowly. Recall that this phenomenon is possible in models that are almost, but not everywhere, identifiable (Remark 14). In these examples LISREL produced similarly divergent sequences with approximately the same likelihood, and 'sem' reported convergence in one case but gave an estimate whose likelihood was nearly 40 points smaller than the intermediate estimates computed by LISREL and RICF.

The columns labelled 'All converge' and 'All agree' in Table 1 show how often all methods converged, and when this occurred, how often the three computed maxima of the log-likelihood function were the same up to one decimal place. Since all methods are for local maximization, substantial disagreements in the computed maxima can occur if the likelihood function is multimodal.

Finally, the last three columns give average CPU time use for the cases in which all three algorithms converge. These are quoted to show that RICF is competitive in terms of computational efficiency, but for the following reasons the precise times should not be used for a formal comparison. On the one hand, LISREL is fast because it is compiled code. This is not the case for the R-based 'sem' and RICF. On the other hand, the fitting routines in LISREL and 'sem' not only compute the MLE but also produce various other derived quantities of interest. This is in contrast to our RICF routine, which only computes the MLE.

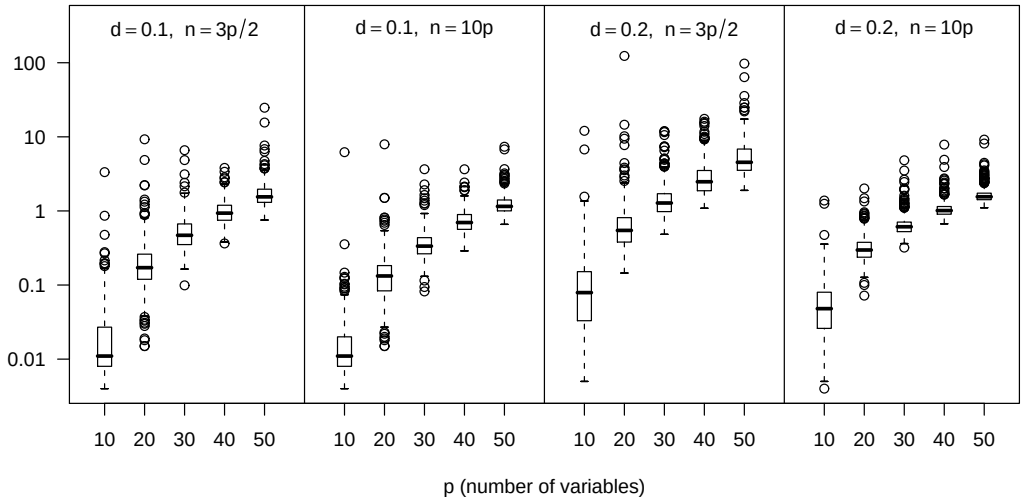

Figure 5: Boxplots of CPU times (in sec. on $\log_{10}$-scale) used by RICF when fitting BAP models to simulated data. Each boxplot summarizes 500 simulations. The number of variables is denoted by p, the sample size is n, and the parameter d determines the expected number of edges of the simulated BAPs (see text for details).

5.2 Simulated Data

In order to demonstrate how RICF behaves when fitting larger models we use the algorithm on simulated data. We consider different choices for the number of variables p and generate random BAPs according to the procedure used in §5.1. We limit ourselves to two different settings for the expected number of edges, choosing $d = 0.1$ or $d = 0.2$ and setting $b = d/2$ in each case. For each BAP G, we simulate a covariance matrix $\Sigma = (I-B)^{-1}\Omega(I-B)^{-t} \in \mathbf{P}(G)$ as follows. The free entries in $B \in \mathbf{B}(G)$ and the free off-diagonal entries in $\Omega \in \mathbf{O}(G)$ are drawn from a $\mathcal{N}(0,1)$ distribution. The diagonal entries ω_{ii} are obtained by adding a draw from a χ^2_1-distribution to the sum of the absolute values of the off-diagonal entries in the i-th row of Ω. This makes Ω diagonally dominant and thus positive definite. Finally, we draw a sample of size n from the resulting multivariate normal distribution $\mathbf{N}(G)$. For each distribution two cases, namely $n = 3p/2$ and $n = 10p$, are considered to illustrate sample size effects. For each combination of p, d and n, we simulate 500 BAPs and associated data sets.

Figure 5 summarizes the results of our simulations in boxplots. As could be expected, the running time for RICF increases with the number of variables p and the expected number of edges in the BAP (determined by d). Moreover, the running time decreases for increased sample size n, which is plausible because the empirical covariance matrix of a larger sample tends to be closer to the underlying parameter space $\mathbf{P}(G)$. The boxplots show that even with $p = 50$ variables the majority of the RICF computations terminate within a few seconds. However, there are also a number of computations in which the running time is considerably longer, though still under two

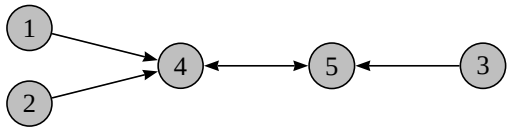

Figure 6: Path diagram for seemingly unrelated regressions.

minutes. This occurs in particular for the denser case with smaller sample size ($d = 0.2$ and $n = 3p/2$).

6. Discussion

As mentioned in the introduction, normal linear models associated with path diagrams are employed in many applied disciplines. The models, also known as structural equation models, have a long tradition but remain of current interest in particular due to the more recent developments in causal inference; compare, for example, Pearl (2000) and Spirtes et al. (2000). Despite their long tradition, however, many mathematical, statistical and computational problems about these models remain open.

The new contribution of this paper is the Residual Iterative Conditional Fitting (RICF) algorithm for maximum likelihood estimation in BAP models. Software for computation of MLEs in structural equation models often employs optimization methods that are not designed to deal with positive definiteness constraints on covariance matrices. This can be seen in particular in Table 1 which shows that two available programs, LISREL (Jöreskog and Sörbom, 1997) and the R package 'sem' (Fox, 2006), fail to converge in a rather large number of problems. This is in line with previous experience by other authors; see, for example, Steiger (2001). Our new RICF algorithm, on the other hand, does not suffer from these problems. It has clear convergence properties (Theorem 13) and can handle problems with several tens of variables (see Figure 5). In addition, RICF has the desirable feature that it estimates parameters in closed form (in a single cycle of its iterations) if this is possible. If applied to a model based on a directed acyclic graph (DAG), the algorithm converges in a single cycle and performs exactly the regressions commonly used for fitting multivariate normal DAG models. This feature and the fact that RICF can be implemented using nothing but least squares computations make it an attractive alternative to less specialized optimization methods.

In another special case, namely seemingly unrelated regressions, RICF reduces to the algorithm of Telser (1964). A path diagram representing seemingly unrelated regressions is shown in Figure 6. The variables Y_1, Y_2 and Y_3 are then commonly thought of as covariates. Since they have no spouses, the MLEs of the variances ω_{11}, ω_{22} and ω_{33} are equal to the empirical variances s_{11}, s_{22} and s_{33}. For the remaining variables Y_i, $i = 4, 5$, RICF performs regressions on both the "covariates" $Y_{\mathrm{pa}(i)}$ and the residual ε_j, $j \in \{4,5\} \setminus \{i\}$. These are precisely the steps performed by Telser.

Existing structural equation modelling software also fits models with latent variables, whereas the RICF algorithm applies only to BAP models without latent variables. However, RICF could be used to implement the M-step in the EM algorithm (Kiiveri, 1987) in order to fit latent variable models. This EM-RICF approach would yield an algorithm with theoretical convergence properties.

Finally, we emphasize that the RICF algorithm is determined by the path diagram. However, different path diagrams may induce the same statistical model; recall point (1) in §1.2 in the intro-

duction. This model equivalence of path diagrams may be exploited to find a diagram for which the running time of RICF is short. For example, for every BAP that is equivalent to a DAG model, parameter estimates could be computed in closed form and hence in finitely many steps. Relevant graphical constructions for this problem are described in Drton and Richardson (2008) and Ali et al. (2005).

Acknowledgments

This work was supported by the U.S. National Science Foundation (DMS-0505612, DMS-0505865, DMS-0746265); the Institute for Mathematics and its Applications; the U.S. National Institutes for Health (R01-HG2362-3, R01-AI032475).

Appendix A. Proofs

Proof [Proof of Proposition 10] Let β and ω be the vectors of unconstrained elements in B and Ω, respectively. The second derivatives of the log-likelihood function with respect to β and ω are:

$$\frac{\partial^2 \ell(B,\Omega)}{\partial\beta\,\partial\beta^t} = -N \cdot P^t \big(S \otimes \Omega^{-1}\big) P, \tag{19}$$

$$\frac{\partial^2 \ell(B,\Omega)}{\partial\beta\,\partial\omega^t} = -N \cdot P^t \big[S(I-B)^t\Omega^{-1} \otimes \Omega^{-1}\big] Q, \tag{20}$$

$$\begin{aligned}\frac{\partial^2 \ell(B,\Omega)}{\partial\omega\,\partial\omega^t} = -\frac{N}{2} Q^t \big\{ &\big[\Omega^{-1} \otimes \Omega^{-1}(I-B)S(I-B)^t\Omega^{-1}\big] \\ &+ \big[\Omega^{-1}(I-B)S(I-B)^t\Omega^{-1} \otimes \Omega^{-1}\big]\big\} Q.\end{aligned} \tag{21}$$

Replacing S by $\mathrm{E}[S] = (I-B)^{-1}\Omega(I-B)^{-t}$ in (19)-(21) yields the claim. ■

Proof [Proof of Proposition 11] If G is a bi-directed chain graph, then the submatrix $B_{\tau_t,\tau_t} = 0$ for all t, while for $s \neq t$ we have $\Omega_{\tau_s,\tau_t} = 0$. In this case the second derivative of the log-likelihood function with respect to β_{ij} and ω_{kl} is equal to $\partial^2\ell(B,\Omega)/\partial\beta_{ij}\,\partial\omega_{kl} = [(I-B)^{-1}]_{jl}\,(\Omega^{-1})_{ik}$. Now $[(I-B)^{-1}]_{jl}$ may only be non-zero if $j = l$ or l is an ancestor of j, that is, if there exists a directed path $l \to j_1 \to \cdots \to j_m \to j$ in G. On the other hand, $(\Omega^{-1})_{ik} = 0$ whenever i and k are not in the same chain component. Therefore, the second derivative in (20) is equal to zero.

Conversely, it follows that the second derivative in (20) vanishes for all parameters only if the graph belongs to the class of bi-directed chain graphs. ■

References

R. A. Ali, T. S. Richardson, P. Spirtes, and J. Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 10–17, Corvallis, Oregon, 2005. AUAI Press.

R. A. Ali, T. S. Richardson, and P. Spirtes. Markov equivalence for ancestral graphs. *Ann. Statist.*, 37:2808–2837, 2009.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25:505–541, 1997.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons Ltd., Chichester, 1978.

K. A. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

C. Brito and J. Pearl. A new identification condition for recursive models with correlated errors. *Struct. Equ. Model.*, 9:459–474, 2002.

S. Chaudhuri, M. Drton, and T. S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199–216, 2007.

M. Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

M. Drton and M. Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist.*, 33:247–257, 2006.

M. Drton and T. S. Richardson. Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika*, 91:383–392, 2004.

M. Drton and T. S. Richardson. Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Mach. Learn. Res.*, 9:893–914, 2008.

C. H. Edwards. *Advanced Calculus of Several Variables*. Dover Publications, New York, 1994.

F. M. Fisher. *The Identification Problem in Econometrics*. McGraw-Hill, New York, 1966.

J. Fox. Structural equation modeling with the sem package in R. *Struct. Equ. Model.*, 13:465–486, 2006.

R. D. Gill and J. M. Robins. Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.*, 29(6):1785–1811, 2001.

K. Jöreskog and D. Sörbom. *LISREL 8: User's Reference Guide*. Scientific Software International, Lincolnwood, IL, 1997.

R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley, New York, 1997.

H. T. Kiiveri. An incomplete data approach to the analysis of covariance structures. *Psychometrika*, 52:539–554, 1987.

J. T. A. Koster. *Linear structural equations and graphical models*. The Fields Institute, Lecture notes, Toronto, 1999.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

R. P. McDonald. What can we learn from the path equations? Identifiability, Constraints, Equivalence. *Psychometrika*, 67:225–249, 2002.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30:962–1030, 2002.

J. M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 349–405. MIT Press, Cambridge, MA, 1999.

J. M. Robins. Causal models for estimating the effects of weight gain on mortality. *International Journal of Obesity*, 32:S15–S41, 2008.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*. AAAI Press, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

J. H. Steiger. Driving fast in reverse. *J. Amer. Statist. Assoc.*, 96:331–338, 2001.

L. G. Telser. Iterative estimation of a set of linear regression equations. *J. Amer. Statist. Assoc.*, 59: 845–862, 1964.

N. Wermuth. Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, 75:963–972, 1980.

N. Wermuth and D. R. Cox. Joint response graphs and separation induced by triangular systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66:687–717, 2004.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester, 1990.

A. Wille, P. Zimmermann, E. Vranova, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in *arabidopsis thaliana*. *Genome Biology*, 5(11):R92, 2004.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5:161–215, 1934.

J. Zhang and P. Spirtes. A characterization of Markov equivalence classes for ancestral graphical models. Technical Report 168, Dept. of Philosophy, Carnegie-Mellon University, 2005.