Advanced Studies in Pure Mathematics 77, 2018
The 50th Anniversary of Gröbner Bases
pp. 35–86

# Algebraic problems in structural equation modeling

**Mathias Drton**

**Abstract.**

The paper gives an overview of recent advances in structural equation modeling. A structural equation model is a multivariate statistical model that is determined by a mixed graph, also known as a path diagram. Our focus is on the covariance matrices of linear structural equation models. In the linear case, each covariance is a rational function of parameters that are associated to the edges and nodes of the graph. We statistically motivate algebraic problems concerning the rational map that parametrizes the covariance matrix. We review combinatorial tools such as the trek rule, projection to ancestral sets, and a graph decomposition due to Jin Tian. Building on these tools, we discuss advances in parameter identification, i.e., the study of (generic) injectivity of the parametrization, and explain recent results on determinantal relations among the covariances. The paper is based on lectures given at the 8th Mathematical Society of Japan Seasonal Institute.

## Contents



Received September 28, 2016.
2010 *Mathematics Subject Classification.* 62H05, 13P25, 05C90.
*Key words and phrases.* algebraic statistics, covariance matrix, Gaussian distribution, graphical model, Gröbner basis, structural equation model.

## Part I. Structural Equation Models

### §1. Motivation

The following example serves well to introduce the statistical models we will consider. It features the simplest instance of what is known as an instrumental variable model. An empirical study that shows this type of model 'in action' can be found in [29].

**Example 1.1.** Does a mother's smoking during pregnancy harm the baby? To answer this question researchers conduct a study in which they record, for a sample of pregnancies, the baby's birth weight and the average number of cigarettes the mom smoked per day during the first trimester. The researchers observe a significant negative correlation between the *birth weight* and *smoking* and are tempted to conclude that smoking has a negative effect on the baby's health, with an increase in the number of cigarettes smoked leading to lower birth weight.

The cigarette companies are not surprised by this finding. They argue, however, that smoking does not harm baby. Instead, heavier smoking merely reflects underlying factors that are the true causes of low birth weight. Such confounding variables could, for instance, be of socio-economic nature. In the context of the smoking-lung cancer debate, prominent Statistician Ronald Fisher liked to argue that correlations can be attributed to unobserved variables of genetic nature [59].

Familiar with this type of counter-argument, the researchers cleverly recorded a third variable: The *tax rate* on tobacco products in the local jurisdictions of the mothers in the sample. It is not unreasonable to assume that the tax rate does not have a direct effect on the baby's health. If there is then variation in the tax rate and higher taxes have an effect on the amount of smoking, then the effect that smoking has on birth weight can be estimated in a model that allows for the presence of unobserved confounders, as we will see shortly.

The above narrative suggests a number of cause-effect relations, as well as the absence thereof. Qualitatively these are summarized in the

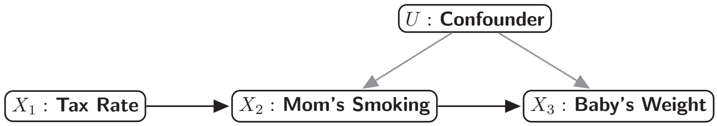


Fig. 1.1. Directed graph for an instrumental variable model.

graph in Figure 1.1. The variables in play are the nodes of the graph and cause-effect relations are indicated as directed edges. Variable $U$ represents an unobserved confounder; we draw its edges in gray.

Structural equation models turn the qualitative descriptions of causes and effects into quantified functional relationships. In this article, the functional relationships will always be *linear*. The linear structural equation model for the present example is based on the following system of *structural equations*:

$$X_1 = \lambda_{01} + \varepsilon_1, \tag{1.1}$$

$$X_2 = \lambda_{02} + \lambda_{12}X_1 + \lambda_{u2}U + \varepsilon_2, \tag{1.2}$$

$$X_3 = \lambda_{03} + \lambda_{23}X_2 + \lambda_{u3}U + \varepsilon_3, \tag{1.3}$$

$$U = \lambda_{0u} + \varepsilon_u. \tag{1.4}$$

Here, the error terms $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$, $\varepsilon_u$ are independent random variables with zero mean. The eight coefficients $\lambda_{01}$, $\lambda_{02}$, $\lambda_{03}$, $\lambda_{0u}$, $\lambda_{12}$, $\lambda_{23}$, $\lambda_{u2}$, and $\lambda_{u3}$ are unknown parameters. Equation (1.1) indicates that variable $X_1$, the tax rate, has expectation $\lambda_{01}$, from which it deviates according to the distribution assumed for $\varepsilon_1$. The analogous statement for the unobserved confounder $U$ is made in (1.4). In (1.2), the amount of smoking, denoted $X_2$, is modeled to be a linear function of the tax rate and independent noise. Similarly, (1.3) introduces birth weight, denoted $X_3$, as a noisy linear function of smoking.

The quantity of primary interest is the coefficient $\lambda_{23}$ that quantifies the relationship between smoking and birth weight. Using data, we can estimate the joint distribution and, in particular, the covariance matrix of the three observed variables $X_1$, $X_2$ and $X_3$. Because the error terms are independent, the covariance between $X_3$ and $X_1$ is

$$\mathrm{Cov}[X_1, X_3] = \lambda_{23}\,\mathrm{Cov}[X_1, X_2]. \tag{1.5}$$

Hence, as long as $\mathrm{Cov}[X_1, X_2] \neq 0$, statistical inference about $\lambda_{23}$ may be based on the ratio of the two covariances in (1.5).

In some applications of structural equation models latent (that is, unobserved) variables are of direct interest. For instance, concepts such as intelligence or depression in psychology are of this nature and measured only indirectly through other variables such as exam results or answers in questionnaires. While problems with explicit latent variables are ubiquitous [5], we will focus on models in which the effects of latent variables are summarized and represented merely in terms of correlations among the error terms in structural equations. This representation of dependence induced by latent variables is discussed in detail in [33, 44, 49, 50, 72].

**Example 1.2.** We take up the instrumental variable model from Example 1.1. The effects of the confounding variable $U$ can be summarized by absorbing $U$ into the error terms in (1.2) and (1.3). Define

$$\tilde{\varepsilon}_2 = \lambda_{u2}U + \varepsilon_2, \qquad \tilde{\varepsilon}_3 = \lambda_{u3}U + \varepsilon_3. \tag{1.6}$$

Retaining only the equations for the observed variables $X_1$, $X_2$, and $X_3$, we are left with the equation system:

$$X_1 = \lambda_{01} + \varepsilon_1, \tag{1.7}$$

$$X_2 = \lambda_{02} + \lambda_{12}X_1 + \tilde{\varepsilon}_2, \tag{1.8}$$

$$X_3 = \lambda_{03} + \lambda_{23}X_2 + \tilde{\varepsilon}_3. \tag{1.9}$$

However, and this is the significance of the unobserved variable $U$, the new error terms may be correlated because

$$\begin{aligned} \omega_{23} &:= \mathrm{Cov}[\tilde{\varepsilon}_2, \tilde{\varepsilon}_3] \\ &= \mathrm{Cov}[\lambda_{u2}U + \varepsilon_2, \lambda_{u3}U + \varepsilon_3] = \lambda_{u2}\lambda_{u3}\,\mathrm{Var}[U] \neq 0. \end{aligned} \tag{1.10}$$

In the sequel, we will focus on models that are given by equations such as (1.7)-(1.9), with one equation for each observed variable but error terms that may be correlated. Graphically, such models may be represented by a *mixed graph* that features directed edges to encode which variables appear in each structural equation and bidirected edges that indicate possibly nonzero correlations between error terms. The mixed graph for the model given by (1.7)-(1.9) is depicted in Figure 1.2, which shows unknown parameters as edge weights. At the nodes, we show the variances of the error terms, namely, $\omega_{11} = \mathrm{Var}[\varepsilon_1]$, $\omega_{22} = \mathrm{Var}[\tilde{\varepsilon}_2]$, and $\omega_{33} = \mathrm{Var}[\tilde{\varepsilon}_3]$. In the statistical literature, the mixed graph for a structural equation model is also known as a *path diagram*.

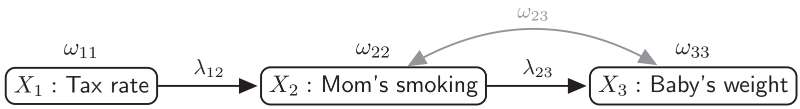


Fig. 1.2. Mixed graph for an instrumental variable model.

The ratio $\mathrm{Cov}[X_2, X_3]/\mathrm{Var}[X_2]$ is the regression coefficient when predicting $X_3$ from $X_2$. We have

$$\text{(1.11)} \qquad \frac{\mathrm{Cov}[X_2, X_3]}{\mathrm{Var}[X_2]} = \lambda_{23} + \frac{\omega_{23}}{\mathrm{Var}[X_2]}.$$

Hence, linear regression predicting $X_3$ from $X_2$ only estimates the coefficient of interest if $\omega_{23} = 0$, as is the case when $X_2$ and $X_3$ do not depend on the latent variable $U$. When $\omega_{23} \neq 0$, the relation from (1.5), which involves all three variables, is needed to recover $\lambda_{23}$.

The remainder of this paper is organized as follows. Section 2 introduces linear structural equation models in full generality. We then formulate questions of statistical interest and the algebraic problems they correspond to (Section 3). Next, we examine the interplay between covariance matrices and mixed graphs. We treat the so-called trek rule (Section 4) and review useful results on subgraphs and graph decomposition (Sections 5 and 6). In Sections 7 and 8, we dive deeper into parameter identifiability, which in this paper means the question of whether a coefficient of interest can be recovered from the covariance matrix of the observed variables. Finally, we discuss relations among the entries of the covariance matrix (Sections 9-12).

## §2. Linear Structural Equation Models

Let $\varepsilon = (\varepsilon_i : i \in V)$ be a random vector indexed by a finite set $V$. Define a new random vector $X = (X_i : i \in V)$ as the solution to the system of so-called structural equations

$$\text{(2.1)} \qquad X = \Lambda^T X + \varepsilon.$$

Here, $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{V \times V}$ is a matrix of unknown parameters. Suppose $\varepsilon$ has covariance matrix $\Omega = (\omega_{ij}) = \mathrm{Var}[\varepsilon]$ with $\Omega$ being a positive definite matrix whose entries are again unknown parameters. Write $I$ for the identity matrix. If $I - \Lambda$ is invertible, then the linear system in (2.1) is solved uniquely by $X = (I - \Lambda)^{-T}\varepsilon$, which has covariance matrix

$$\mathrm{Var}[X] \;=\; (I-\Lambda)^{-T}\Omega(I-\Lambda)^{-1} \;\;=:\;\; \phi(\Lambda,\Omega). \tag{2.2}$$

Interesting settings are obtained by restricting the support of $\Lambda$ and $\Omega$, as is the case in our motivating example.

**Example 2.1.** Consider the setup from Example 1.2. If the equation system from (1.7)-(1.9) is written in vector form as in (2.1), then the coefficient matrix is

$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & 0 \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.3}$$

The error covariance matrix is

$$\Omega = \mathrm{Var}[\varepsilon] = \begin{pmatrix} \omega_{11} & 0 & 0 \\ 0 & \omega_{22} & \omega_{23} \\ 0 & \omega_{23} & \omega_{33} \end{pmatrix}. \tag{2.4}$$

From (2.2), the covariance matrix of $X = (X_1, X_2, X_3)$ is found to be

$$\mathrm{Var}[X] \;=\; \begin{pmatrix} \omega_{11} & \omega_{11}\lambda_{12} & \omega_{11}\lambda_{12}\lambda_{23} \\ \omega_{11}\lambda_{12} & \omega_{22}+\omega_{11}\lambda_{12}^2 & \omega_{23}+\lambda_{23}\sigma_{22} \\ \omega_{11}\lambda_{12}\lambda_{23} & \omega_{23}+\lambda_{23}\sigma_{22} & \omega_{33}+2\omega_{23}\lambda_{23}+\lambda_{23}^2\sigma_{22} \end{pmatrix}, \tag{2.5}$$

where $\sigma_{22}$ denotes the $(2,2)$ entry of $\mathrm{Var}[X]$. The relation from (1.5) can be confirmed from the $(1,2)$ and $(2,3)$ entry of $\mathrm{Var}[X]$.

Restrictions on the support of a matrix naturally correspond to a graph. Specifically, we adopt mixed graphs because we are dealing with two matrices, $\Lambda$ and $\Omega$, whose rows and columns are indexed by the same set $V$. In structural equation modeling, this point of view originated in the work of Sewall Wright [73, 74].

A mixed graph with vertex set $V$ is a triple $G = (V, D, B)$ where $D$ and $B$ are two sets of edges. The set $D \subset V \times V$ contains ordered pairs $(i,j)$, which we also denote by $i \to j$ to visualize that such a pair encodes a directed edge pointing from $i$ to $j$. Then $i$ is the tail and $j$ is the head of the edge. The elements of $B$ are unordered pairs $\{i,j\}$ and encode bidirected edges that we also denote by $i \leftrightarrow j$. These edges have no orientation, so $i \leftrightarrow j \in B$ if and only if $j \leftrightarrow i \in B$. It will be convenient to call both endpoints $i$ and $j$ heads of $i \leftrightarrow j$. In our context, neither the bidirected part $(V,B)$ nor the directed part $(V,D)$ contain loops, so $i \to i \notin D$ and $i \leftrightarrow i \notin B$ for all $i \in V$. The mixed graph $G$ is acyclic if $(V, D)$ does not have any directed cycles $i \to \ldots \to i$.

Let $\mathbb{R}^D$ be the set of real $V \times V$-matrices $\Lambda = (\lambda_{ij})$ with support in $D$, that is,

$$\mathbb{R}^D = \left\{ \Lambda \in \mathbb{R}^{V \times V} : \lambda_{ij} = 0 \ \text{ if } \ i \to j \notin D \right\}. \tag{2.6}$$

Define $\mathbb{R}^D_{\mathrm{reg}}$ to be the subset of matrices $\Lambda \in \mathbb{R}^D$ for which $I - \Lambda$ is invertible. If $G$ is acyclic, then there is a permutation of $V$ that makes $I - \Lambda$ unit upper triangular such that $\det(I - \Lambda) = 1$ for all $\Lambda \in \mathbb{R}^D$ and, thus, $\mathbb{R}^D = \mathbb{R}^D_{\mathrm{reg}}$. Similarly, let $PD_V$ be the cone of positive definite symmetric $V \times V$-matrices $\Omega = (\omega_{ij})$, and define $PD(B)$ to be the subcone of matrices supported over $B$, that is,

$$PD(B) = \left\{ \Omega \in PD_V : \omega_{ij} = 0 \ \text{ if } \ i \neq j \text{ and } i \leftrightarrow j \notin B \right\}. \tag{2.7}$$

Taking the error vector $\varepsilon$ to be Gaussian (or in other words, to follow a multivariate normal distribution), we arrive at the following definition of a statistical model for the random vector $X$ that solves (2.1). Readers looking for background such as the fact that linear transformations of a Gaussian random vector are Gaussian may consult a textbook on multivariate statistics, e.g., [2].

**Definition 2.2.** *The* linear structural equation model *given by a mixed graph $G = (V, D, B)$ is the family of all multivariate normal distributions on $\mathbb{R}^V$ with covariance matrix in the set*

$$\mathcal{M}_G \;=\; \left\{ (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} : \Lambda \in \mathbb{R}^D_{\mathrm{reg}},\, \Omega \in PD(B) \right\}.$$

*The* covariance parametrization *of the model is the map*

$$\phi_G : \mathbb{R}^D \times PD(B) \to PD_V, \qquad (\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

*The* fiber *of a pair $(\Lambda, \Omega) \in \mathbb{R}^D_{\mathrm{reg}} \times PD(B)$ is the preimage*

$$\mathcal{F}_G(\Lambda, \Omega) \;=\; \left\{ (\Lambda', \Omega') \in \mathbb{R}^D_{\mathrm{reg}} \times PD(B) \;:\; \phi_G(\Lambda', \Omega') = \phi_G(\Lambda, \Omega) \right\}.$$

As defined, a linear structural equation model does not impose any restrictions on the mean vector of the normal distributions. Consequently, the mean vector plays no role in our discussion. For instance, in maximum likelihood estimation we may assume without loss of generality that the mean vector is zero. Other questions we consider will directly concern covariances. Therefore, we may safely identify a linear structural equation model with its set of covariance matrices $\mathcal{M}_G$. On occasion, we will simply refer to $\mathcal{M}_G$ as the model.

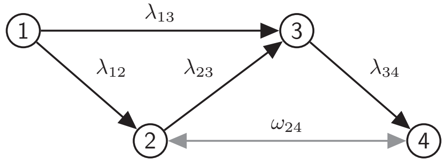


Fig. 2.1. An acyclic mixed graph known as the Verma graph.

Leaving statistics out of the picture, our interest is in the maps $\phi_G$, their fibers $\mathcal{F}_G$ and their images $\mathcal{M}_G$. Algebra comes into play naturally.

**Proposition 2.3.** *For any mixed graph $G$, the map $\phi_G$ is a rational map whose image $\mathcal{M}_G$ and fibers $\mathcal{F}_G(\Lambda, \Omega)$ are semi-algebraic sets. The map $\phi_G$ is a polynomial map if and only if $G$ is acyclic.*

*Proof.* That $\phi_G$ is rational follows from Cramer's rule for matrix inversion. The domain of $\phi_G$ is a semi-algebraic set and, thus, the fibers $\mathcal{F}_G(\Lambda, \Omega)$ are semi-algebraic as well. The Tarski-Seidenberg theorem implies that $\mathcal{M}_G$ is semi-algebraic. If $G = (V, D, B)$ is acyclic, then $\det(I - \Lambda) = 1$ for all $\Lambda \in \mathbb{R}^D$. Consequently, the entries of $(I - \Lambda)^{-1}$ are polynomial in $\Lambda$. If $G$ is not acyclic, then $\det(I - \Lambda)$ is a non-constant polynomial. By the Leibniz formula, its terms correspond to collections of disjoint directed cycles in the graph; compare Theorem 1 in [41]. Q.E.D.

**Example 2.4.** The mixed graph $G = (V, D, B)$ in Figure 2.1 encodes the structural equations

$$\begin{array}{lclr} X_1 & = & \lambda_{01} & +\varepsilon_1, \\ X_2 & = & \lambda_{02} + \lambda_{12} X_1 & +\varepsilon_2 \\ X_3 & = & \lambda_{03} + \lambda_{13} X_1 + \lambda_{23} X_2 & +\varepsilon_3 \\ X_4 & = & \lambda_{04} + \lambda_{34} X_3 & +\varepsilon_4. \end{array}$$

Only the errors $\varepsilon_2$ and $\varepsilon_4$ may be dependent and the error covariance matrix is

$$\Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & \omega_{24} \\ 0 & 0 & \omega_{33} & 0 \\ 0 & \omega_{24} & 0 & \omega_{44} \end{pmatrix}.$$

Subtracting the coefficient matrix from the identity gives

$$I - \Lambda = \begin{pmatrix} 1 & -\lambda_{12} & -\lambda_{13} & 0 \\ 0 & 1 & -\lambda_{23} & 0 \\ 0 & 0 & 1 & -\lambda_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with $\det(I-\Lambda) = 1$ and inverse

$$(I-\Lambda)^{-1} = \begin{pmatrix} 1 & \lambda_{12} & \lambda_{13}+\lambda_{12}\lambda_{23} & \lambda_{13}\lambda_{34}+\lambda_{12}\lambda_{23}\lambda_{34} \\ 0 & 1 & \lambda_{23} & \lambda_{23}\lambda_{34} \\ 0 & 0 & 1 & \lambda_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

To illustrate the form of the map $\phi_G$, we display the coordinate function

$$(2.8)\quad \phi_G(\Lambda,\Omega)_{24} \;=\; \lambda_{12}\lambda_{13}\lambda_{34}\omega_{11} + \lambda_{12}^2\lambda_{23}\lambda_{34}\omega_{11} + \lambda_{23}\lambda_{34}\omega_{22} + \omega_{24}.$$

## §3. Questions of Interest

Structural equation models are used to empirically estimate, test and possibly discover cause-effect relationships among a set of variables. In estimation and testing, the underlying graph is given. In discovery, we seek to infer the underlying graph, or in other words, perform model selection. This section gives a broad overview of algebraic problems that arise in the context of these statistical tasks. Only some of the problems are treated in the remainder of the paper, which focuses on parameter identifiability and polynomial relations between covariances.

### 3.1. Parameter identification

When specifying a model via a mixed graph $G = (V, D, B)$, a first question is whether the effects of interest are *identifiable*, that is, whether they are determined by the joint distribution of the observed variables. The importance of the question is clear: The joint distribution is what can be estimated from data. For linear and Gaussian models, the problem is equivalent to deciding whether the coefficients $\lambda_{ij}$ in the structural equations can be recovered from the covariance matrix of the variables.

Different notions of parameter identifiability translate into related but different algebraic problems. The most stringent identifiability property is for a model to have all of its coefficients $\lambda_{ij}$, $i \to j \in D$, identifiable. In this case, we seek to answer the following:

**Question 3.1.** *Is the map $\phi_G$ is injective?*

Injectivity of $\phi_G$ can be decided efficiently, as we will discuss in Section 7. However, injectivity can be too strong of a requirement because all fibers are required to be singletons with $\mathcal{F}_G(\Lambda,\Omega) = \{(\Lambda,\Omega)\}$. Indeed, some interesting examples have fibers that are not singletons.

**Example 3.2.** The map $\phi_G$ fails to be injective when $G$ is the graph for the instrumental variable model from Example 1.2. The relation from (1.5) shows that $\mathcal{F}_G(\Lambda,\Omega) = \{(\Lambda,\Omega)\}$ if $\lambda_{12} \neq 0$. If $\lambda_{12} = 0$, however, then the fiber is infinite. Hence, all model parameters are identifiable as long as $\lambda_{12} \neq 0$. In the context of Example 1.2, this requires making an argument that higher tax rates impact the amount of smoking.

In the example just given, $\mathcal{F}_G(\Lambda, \Omega) = \{(\Lambda, \Omega)\}$ for generic choices of $(\Lambda, \Omega) \in \mathbb{R}^D_{\text{reg}} \times PD(B)$. In this case, we call $\phi_G$ *generically injective*. We are led to:

**Question 3.3.** *Is the map $\phi_G$ is generically injective?*

Generic injectivity turns out to be more difficult to decide. The problem's computational complexity has not yet been determined. In Section 8, we review methods to decide whether $\phi_G$ is generically injective as well as methods to decide when the fibers are generically infinite.

When $\phi_G$ is not generically injective, its generic fibers may be discrete sets. This property is known as *local identifiability* in the statistical literature. We will instead speak of $\phi_G$ being *generically finite-to-one* to highlight that in our case a discrete fiber is in fact finite because $\phi_G$ is rational. By the inverse function theorem, the question of whether $\phi_G$ is generically finite-to-one is the same as:

**Question 3.4.** *Does the Jacobian of $\phi_G$ have full column rank?*

The fiber $\mathcal{F}_G(\Lambda, \Omega)$ is defined by the equation system $\phi_G(\Lambda', \Omega') = \phi_G(\Lambda, \Omega)$. These equation systems have a generic number of complex solutions (i.e., the free entries of $\Lambda$ and $\Omega$ are allowed to be complex numbers).

**Definition 3.5.** *The map $\phi_G$ is algebraically $k$-to-one if the equation systems defining its fibers generically have $k$ complex solutions. We call the number $k$ the algebraic degree of identifiability of $G$.*

The degree of identifiability may be determined by Gröbner basis methods (see Section 8). It is finite if and only if $\phi_G$ is generically finite-to-one. The main theorem in Section 7.6 shows that if $\phi_G$ is injective then its inverse is rational, which is the same as $G$ having degree of identifiability one. Currently, there are no combinatorial results for when the degree is finite but larger than one. Example 8(b) in [34] has degree 3 but fibers whose cardinality over the reals is either one or three. To the author's knowledge, no example has been discovered in which $\phi_G$ is generically injective over the reals but algebraically $k$-to-one for $k \geq 2$.

An important question that we will not address in detail is the identifiability of only a single parameter $\lambda_{ij}$ for a designated edge of interest $i \to j \in D$. This amounts to checking whether in every fiber the coefficient for the edge has only a single value. In other words, it must hold that $\lambda'_{ij} = \lambda''_{ij}$ whenever $(\Lambda', \Omega')$ and $(\Lambda'', \Omega'')$ are in the same fiber. In Example 1.2, the fiber of a pair $(\Lambda, \Omega)$ with $\lambda_{12} = 0$ is infinite but all pairs $(\Lambda', \Omega')$ such a fiber have $\lambda'_{12} = 0$. Two well-known graphical methods for identifying a single edge coefficient are the back-door and the front-door criterion [49]; see also [11, 36].

### 3.2. Model dimension

Statistical tests may be used to assess whether a model is compatible with empirical data. At an intuitive level, such tests are based on computing a distance between data and model and comparing this distance to typical distances that are obtained when data are generated from a distribution in the model. For linear Gaussian models, a test can be thought off as assessing the distance between the empirical (or sample) covariance matrix and the model $\mathcal{M}_G$. Recall that we have defined $\mathcal{M}_G$ as the set of covariance matrices.

The challenging part of designing a statistical test is to quantify, in a probabilistic manner, what typical distances between data and model are. Many procedures rely on asymptotic approximations that are obtained by letting the number of data points grow to infinity. Under regularity conditions, limiting distribution theory leads to consideration of so-called chi-square distributions, which are indexed by an integer parameter. In our context, when testing the model given by the mixed graph $G = (V, D, B)$, the chi-square parameter is set equal to the codimension of $\mathcal{M}_G$, where we think of $\mathcal{M}_G$ as embedded in the space of symmetric matrices. This gives concrete statistical motivation for:

**Question 3.6.** *What is the dimension of $\mathcal{M}_G$?*

The model $\mathcal{M}_G$ is parametrized by the coefficients and covariances associated with the edges in $D$ and $B$ as well as the variances associated with the nodes in $V$. Being a subset of the space of $V \times V$ symmetric matrices, $\mathcal{M}_G$ has expected dimension

$$\min\left\{|V| + |D| + |B|, \frac{|V|(|V|+1)}{2}\right\}.$$

The term $|D|+|V|+|B|$ counts the nodes and edges of $G$. Since $\mathcal{M}_G$ is the image of $\phi_G$, its actual dimension is equal to the rank of the Jacobian of $\phi_G$. A review of the connection between dimension and Jacobian in a statistical context is given in [38]. Question 3.6 is tied to parameter identifiability, most directly to Question 3.4. If $\phi_G$ is generically finite-to-one, then $\mathcal{M}_G$ has the expected dimension $|V| + |D| + |B|$.

### 3.3. Covariance equivalence

Different graphs may induce the same statistical model. For example, take $V = \{1, 2\}$, and let $G_1$ be the graph with the single edge $1 \to 2$. Let $G_2$ and $G_3$ be the graphs with single edge $1 \leftarrow 2$ and $1 \leftrightarrow 2$, respectively. Then $\mathcal{M}_{G_1} = \mathcal{M}_{G_2} = \mathcal{M}_{G_3}$ as each model is easily seen to be equal to the entire cone of positive definite $2 \times 2$ matrices.

From an applied perspective, two different graphs $G$ and $G'$ encode different scientific/causal hypotheses. If $\mathcal{M}_G = \mathcal{M}_{G'}$, then the two hypotheses cannot be distinguished based on data from a linear and Gaussian structural equation model. It is thus useful to be able to decide whether two graphs $G$ and $G'$ are covariance equivalent, that is, we would like to be able to answer:

**Question 3.7.** *When do two maps $\phi_G$ and $\phi_{G'}$ have the same image?*

Existing results addressing this questions make comparisons between certain types of relations among the entries of the covariance matrices in each model. This ties into the basic problem of implicitization:

**Question 3.8.** *What are the algebraic relations among the coordinates of $\phi_G$?*

Such relations are also of interest for statistical tests that assess whether the model given by $G$ is compatible with available data [6, 11, 17]. We review results on relations among the covariances in Sections 9-12. An important role is played by determinants that represent probabilistic conditional independence in a Gaussian random vector. We note that models can, in principle, also be distinguished using inequality constraints. However, as less is know about inequalities, we do not treat them here. Examples of models with latent variables for which a full semi-algebraic description is available can be found in [25, 53].

*Remark* 3.9. As defined above, covariance equivalence is based on data that is observational, i.e., it is collected by merely observing the considered physical system. The situation is different when experimental data is available, i.e., (some) data is collected in settings in which the system is subject to various experimental interventions. We will not treat such interventional data in this paper. Interested readers may find discussions of the problem in [43, 49, 57]. Similarly, even for observational data, questions of equivalence differ from Question 3.7 in non-linear models or linear models with non-Gaussian errors [26, 54].

### 3.4. Maximum likelihood

The parameters of linear structural equation models are most commonly estimated using the technique of maximum likelihood. Suppose we observe a sample $X^{(1)}, \ldots, X^{(n)}$ drawn independently from the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu$ is the mean vector and $\Sigma$ is the covariance matrix. The joint distribution of the random vectors $X^{(1)}, \ldots, X^{(n)}$ is the $n$-fold product of $\mathcal{N}(\mu, \Sigma)$. The likelihood of the sample is the value of the joint density of the product distribution at $(X^{(1)}, \ldots, X^{(n)})$. The *likelihood function* maps the pair $(\mu, \Sigma)$ to the likelihood of the sample. The *maximum likelihood estimator* (MLE) of $(\mu, \Sigma)$ under the model given by a mixed graph $G$ is the maximizer of the likelihood function when restricting $\Sigma$ to be in $\mathcal{M}_G$.

Define the sample mean vector and the sample covariance matrix as

$$(3.1) \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} \quad \text{and} \quad S_n = \frac{1}{n}(X^{(i)} - \bar{X}_n)(X^{(i)} - \bar{X}_n)^T,$$

respectively. It is convenient to treat the likelihood function on the log-scale. With an additive constant omitted and $n/2$ divided out, the *log-likelihood function* is

$$(\mu, \Sigma) \mapsto -\log\det(\Sigma) - \operatorname{trace}\left(\Sigma^{-1} S_n\right) - (\bar{X}_n - \mu)^T \Sigma^{-1} (\bar{X}_n - \mu).$$

Because the considered models place no constraint on the mean vector, its MLE is always $\bar{X}_n$. The MLE of $\Sigma$ maximizes the function

$$(3.2) \qquad \ell(\Sigma) = -\log\det(\Sigma) - \operatorname{trace}\left(\Sigma^{-1} S_n\right)$$

subject to $\Sigma \in \mathcal{M}_G$. Using the covariance parametrization, the MLE is found by maximizing $\ell \circ \phi_G$. A key problem is then understanding the existence and uniqueness of the MLE. We record:

**Question 3.10.** *For which sample covariance matrices $S_n$ does the likelihood function $\ell \circ \phi_G$ achieve its maximum?*

Graphical models theory solves Question 3.10 when $G = (V, D, B)$ is an acyclic digraph, i.e., has $B = \emptyset$ and $D$ without directed cycles [46]. More generally, it is well known that $\ell \circ \phi_G$ is bounded when $S_n$ is positive definite but this is not necessary [32]. An issue that is not well explored is the fact that even if $\ell \circ \phi_G$ is bounded it may fail to achieve its maximum as the model $\mathcal{M}_G$ need not be closed. For instance, the model in Example 3.14 is not closed. We remark that Question 3.10 is closely related to a positive definite matrix completion problem that arises in ML estimation for other types of graphical models [9, 63, 67].


Fig. 3.1. Mixed graph for a bivariate seemingly unrelated regressions model.

In some models, the MLE is known to admit a closed-form expression as a rational function of the data. Such models have *maximum likelihood* (*ML*) *degree* equal to one, in the sense of the following:

**Question 3.11.** *The MLE of $\Sigma$ in model $\mathcal{M}_G$ is an algebraic function of the data. What is the degree of this function?*

An introduction to the notion of ML degree is given in [21, Chapter 2]. Here, we merely not that the ML degree is one when $G$ is an acyclic digraph. More general models may have higher ML degree and a log-likelihood function with more than one local maximum. We exemplify this for a model discussed in detail in [19].

**Example 3.12.** Suppose the sample covariance matrix is

$$S_n = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} 8 & -5 & 10 & 3 \\ -5 & 27 & 4 & 49 \\ 10 & 4 & 21 & 24 \\ 3 & 49 & 24 & 114 \end{bmatrix} \end{array}.$$

The matrix is positive definite such that the log-likelihood function $\ell$ from (3.2) is bounded above on the entire cone of positive definite matrices. Moreover, $\ell$ has compact level sets, i.e., for any constant $c \in \mathbb{R}$ the set of positive definite matrices $\Sigma$ with $\ell(\Sigma) \geq c$ is compact [2].

Let $G$ be the graph from Figure 3.1. It is not difficult to show that the parametrization $\phi_G$ admits a rational inverse. Let $\Sigma = (\sigma_{ij})$ satisfy $\Sigma = \phi_G(\Lambda, \Omega)$ with $\Lambda = (\lambda_{ij}) \in \mathbb{R}^D$ and $\Omega = (\omega_{ij}) \in PD(B)$. Then

$$\lambda_{12} = \frac{\sigma_{12}}{\sigma_{11}}, \qquad \lambda_{43} = \frac{\sigma_{34}}{\sigma_{44}}, \tag{3.3}$$

and the entries of $\Omega = (I - \Lambda)^T \Sigma (I - \Lambda)$ are rational functions of $\Sigma$ as well. All the rational functions are defined on the entire cone of positive definite matrices because $\sigma_{11}, \sigma_{44} > 0$. It is also clear that the considered map $\phi_G$ is proper, that is, compact subsets of the positive definite cone have compact preimages under $\phi_G$. It follows that $\ell \circ \phi_G$ has compact level sets and, thus, achieves its maximum on the open set $\mathbb{R}^D \times PD(B)$.

The critical points of $\ell \circ \phi_G$ satisfy a rational equation system, in which the determinant of $\phi_G(\Lambda, \Omega)$ appears in the denominator. Since $G$ is acyclic the determinant is equal to the determinant of $\Omega$. Clearing the denominator yields a polynomial equation system. Saturating the system with respect to the determinant removes infeasible solutions with $\det(\Omega) = 0$. Computing a lexicographic Gröbner basis after the saturation shows that the critical points $(\Lambda, \Omega)$ solve the equation

$$\begin{aligned} 10583160\,\lambda_{12}^5 + 43115307\,\lambda_{12}^4 + 72738452\,\lambda_{12}^3 \\ + 55482894\,\lambda_{12}^2 + 8437660\,\lambda_{12} - 4703765 \;=\; 0. \end{aligned}$$

All other entries of $\Lambda$ and also $\Omega$ solve linear equations whose coefficients depend on $\lambda_{12}$ and the data. We conclude that the MLE of $\Sigma$ is an algebraic function of degree 5. The displayed equation for $\lambda_{12}$ has three real roots and, thus, is not solvable by radicals.

### 3.5. Model singularities

As noted in Section 3.2, the distributions of test statistics are frequently approximated using asymptotic theory. For so-called likelihood ratio tests, this asymptotic theory can be thought of as assessing infinitesimal distances between a positive semidefinite data matrix and the given model $\mathcal{M}_G$. The data matrix is generated from a distribution that corresponds to a particular point in $\mathcal{M}_G$. At a smooth point of $\mathcal{M}_G$, the squared infinitesimal distance follows a chi-square distribution, which is the distribution of the squared Euclidean distance between a Gaussian random vector and a linear space. At singular points, the distribution is determined by the tangent cone at the considered point [15]. Singularities also impact other approaches such as Wald tests [24], and it is important to clarify:

**Question 3.13.** *Is the image of $\phi_G$ a smooth manifold? If not, what are the tangent cones of the image?*

We do not address the question explicitly in this paper. However, whenever $\phi_G$ is injective (see Section 7) its image is smooth. Indeed, when $\phi_G$ is injective it has a rational inverse whose domain of definition includes the cone of all positive definite matrices [16]. The next example illustrates that not all models are smooth.

**Example 3.14.** Consider the mixed graph $G = (V, D, B)$ from Figure 3.2. Let $\Sigma \in \mathbb{R}^{4\times 4}$ be a positive definite matrix. Define the matrix

$$\Sigma_{\{3,4\}.\{1,2\}} = \Sigma_{\{3,4\},\{1,2\}} \left(\Sigma_{\{1,2\},\{1,2\}}\right)^{-1},$$

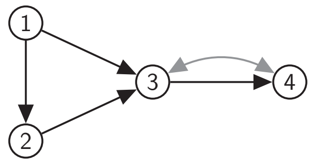


Fig. 3.2. Mixed graph of a model with two instruments.

and the Schur complement

$$\Sigma_{\{3,4\},\{3,4\}.\{1,2\}} = \Sigma_{\{3,4\},\{3,4\}} - \Sigma_{\{3,4\},\{1,2\}} \left(\Sigma_{\{1,2\},\{1,2\}}\right)^{-1} \Sigma_{\{1,2\},\{3,4\}}.$$

Change coordinates to the triple of $2 \times 2$ matrices

$$\left(\Sigma_{\{1,2\},\{1,2\}},\ \Sigma_{\{3,4\}.\{1,2\}},\ \Sigma_{\{3,4\},\{3,4\}.\{1,2\}}\right).$$

If $\Sigma = \phi_G(\Lambda, \Omega)$ for $\Lambda = (\lambda_{ij}) \in \mathbb{R}^D$ and $\Omega = (\omega_{ij}) \in PD(B)$, then

$$\begin{aligned}
\Sigma_{\{1,2\},\{1,2\}} &= \begin{pmatrix} \omega_{11} & \lambda_{12}\omega_{11} \\ \lambda_{12}\omega_{11} & \omega_{22} \end{pmatrix}, \\
\Sigma_{\{3,4\}.\{1,2\}} &= \begin{pmatrix} \lambda_{13} & \lambda_{23} \\ \lambda_{13}\lambda_{34} & \lambda_{23}\lambda_{34} \end{pmatrix}, \\
\Sigma_{\{3,4\},\{3,4\}.\{1,2\}} &= \begin{pmatrix} \omega_{33} & \omega_{34} + \lambda_{34}\omega_{33} \\ \omega_{34} + \lambda_{34}\omega_{33} & \omega_{44} + 2\lambda_{34}\omega_{34} + \lambda_{34}^4\omega_{33} \end{pmatrix}.
\end{aligned}$$

We observe that $\Sigma$ is in the (topological) closure of $\mathcal{M}_G$ if and only if $\Sigma_{\{3,4\}.\{1,2\}}$ is a matrix of rank at most one. For $\Sigma$ to be in $\mathcal{M}_G$, the second row of $\Sigma_{\{3,4\}.\{1,2\}}$ may be zero only if the first row is zero.

Geometrically, the closure of $\mathcal{M}_G$ is equivalent to the product of two cones of positive definite $2 \times 2$ matrices and the set of $2 \times 2$ matrices of rank at most one. The latter set is singular at the zero matrix. For more details see the related example in [21, Exercise 6.4].

### 3.6. Singularities of fibers

Finally, without going into any detail, we note that it is also of statistical interest to study the geometry of the fibers $\mathcal{F}_G(\Lambda, \Omega)$. In particular, the resolution of singularities of $\mathcal{F}_G(\Lambda, \Omega)$ is connected to asymptotic approximations in Bayesian approaches to model selection.

Bayesian methods assess the goodness-of-fit of a model by integrating the likelihood function with respect to a prior distribution. In models with many parameters, the integration is over a domain of larger dimension and may constitute a difficult numerical problem. While carefully tuned Monte Carlo methods can be effective, it can also be useful to invoke asymptotics. For large sample size $n$, the integrated likelihood function behaves like a Laplace integral. Under regularity conditions, a Laplace approximation can yield accurate approximations that have been used in many applications [42]. However, the models considered here may also lead to singular Laplace integrals for which asymptotic approximations are more involved.

Asymptotic expansions for singular Laplace integrals are well-studied [3]. The work of Sumio Watanabe brings the ideas to bear in the statistical context [70]. For several practically relevant settings, it has become tractable to determine or bound the *real log-canonical threshold* and its multiplicity, which determine how the integrated likelihood scales with the sample size $n$. This information can be used in model selection [18]. Computing real log-canonical thresholds for data generated under the distribution with covariance matrix $\phi_G(\Lambda, \Omega)$ requires careful study of the singularities of the fiber $\mathcal{F}_G(\Lambda, \Omega)$. Bounds on the thresholds can be obtained from cruder information such as the dimension of the fiber.

## Part II. Treks, Subgraphs and Decomposition

### §4. Trek Rule

In solving the problems from Section 3, it is desirable to exploit the connection between the covariance parametrization $\phi_G$ of a structural equation model and the underlying mixed graph $G = (V, D, B)$. The trek rule that we present in this section makes the connection precise and is behind results that allow one to answer some of the questions we posed with efficient algorithms.

It is natural to expect the covariance between random variables $X_i$ and $X_j$ to be determined by the semi-walks between the nodes $i$ and $j$ in the graph $G$. A semi-walk is an alternating sequence of nodes from $V$ and edges from either $D$ or $B$ such that the endpoints of each edge are the nodes immediately preceding and succeeding the edge in the sequence. In other words, a semi-walk is a walk that uses bidirected or directed edges but is allowed to traverse directed edges in the 'wrong direction'. As we will see, only special semi-walks contribute to the covariance.

**Definition 4.1.** *A* trek *$\tau$ from initial node $i$ to target node $j$ is a semi-walk from $i$ to $j$ whose consecutive edges do not have any colliding arrowheads. In other words, $\tau$ is a sequence of the form*

(a) $i \leftarrow i_l \leftarrow \cdots \leftarrow i_1 \leftarrow i_0 \longleftrightarrow j_0 \to j_1 \to \cdots \to j_r \to j$, *or*

(b) $i \leftarrow i_l \leftarrow \cdots \leftarrow i_1 \longleftarrow i_0 \longrightarrow j_1 \to \cdots \to j_r \to j$.

*A trek has a* left- *and a* right-hand side, *denoted left*$(\tau)$ *and right*$(\tau)$, *respectively. We have left*$(\tau) = \{i_0, \ldots, i_l, i\}$ *and right*$(\tau) = \{j_0, \ldots, j_r, j\}$ *in case* (a), *and left*$(\tau) = \{i_0, \ldots, i_l, i\}$ *and right*$(\tau) = \{i_0, j_1, \ldots, j_r, j\}$ *in case* (b). *In case* (b) *the* top *node* $i_0$ *belongs to both sides.*

In an acyclic graph, if we think of directed edges pointing 'downward', then a trek takes us up and/or down a 'mountain'. A trek $\tau$ from $i$ to $i$ may have no edges, in which case $i$ is the top node, $\mathrm{left}\,(\tau) = \mathrm{right}\,(\tau) = \{i\}$. We call such a trek trivial. Any directed path is a trek, in which case $|\mathrm{left}\,(\tau)\,| = 1$ or $|\mathrm{right}\,(\tau)\,| = 1$ depending on the direction in which the path is traversed. A trek may contain the same node on both its left- and right-hand sides. If the graph contains a cycle, then the left- or right-hand side of $\tau$ may contain this cycle, possibly repeated.

For a trek $\tau$ that contains no bidirected edge and has top node $i_0$, define a trek monomial as

$$\tau(\Lambda, \Omega) = \omega_{i_0 i_0} \prod_{k \to l \in \tau} \lambda_{kl}.$$

For a trek $\tau$ that contains a bidirected edge $i_0 \leftrightarrow j_0$, define the trek monomial as

$$\tau(\Lambda, \Omega) = \omega_{i_0 j_0} \prod_{k \to l \in \tau} \lambda_{kl}.$$

The following rule expresses the covariance matrix as a summation over treks [57, 73, 74]. We write $\mathcal{T}(i,j)$ for the set of all treks from $i$ to $j$.

**Theorem 4.2** (Trek rule)**.** *Let* $G = (V, D, B)$ *be any mixed graph, and let* $\Lambda \in \mathbb{R}^D$ *and* $\Omega \in PD(B)$. *Then the covariances are*

$$\phi_G(\Lambda, \Omega)_{ij} = \sum_{\tau \in \mathcal{T}(i,j)} \tau(\Lambda, \Omega), \qquad i, j \in V. \tag{4.1}$$

Some clarification is in order. If $G$ is acyclic, then the summation in (4.1) is finite and yields a polynomial. If $G$ contains a directed cycle, then the right-hand side of (4.1) may yield a power series as shown in Example 4.4 below. Under suitable assumptions on the spectrum of $\Lambda$, the power series converges and yields the value of $\phi_G(\Lambda, \Omega)_{ij}$. Such spectral conditions are also needed to give cyclic models an interpretation of representing observation of an equilibrium [31, 45]. This said, it is also useful to treat the right-hand side of (4.1) as a formal power series. If so desired, a combinatorial description can also be given for a rational expression for $\phi_G(\Lambda, \Omega)_{ij}$; see [14].

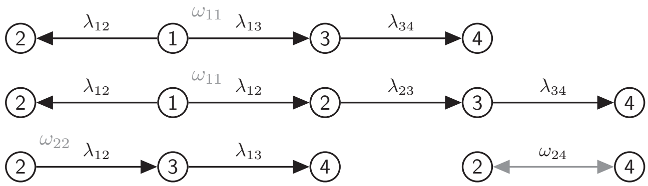


Fig. 4.1. Four treks in the graph from Figure 2.1.

*Proof of the trek rule.* Writing $(I-\Lambda)^{-1} = I + \Lambda + \Lambda^2 + \dots$, we observe that

$$((I-\Lambda)^{-1})_{ij} = \sum_{\tau \in \mathcal{P}(i,j)} \prod_{k \to l \in \tau} \lambda_{kl}, \tag{4.2}$$

where $\mathcal{P}(i,j)$ is the set of directed paths from $i$ to $j$ in $G$. If $G$ is acyclic, then $\Lambda^m = 0$ for all $m \geq |V|$, and the geometric series of matrices has only finitely many nonzero terms. If $G$ is cyclic the geometric series is infinite and converges if and only if all eigenvalues of $\Lambda$ have magnitude less than 1. Now, observe that a product of three entries of $(I-\Lambda)^{-T}$, $\Omega$, and $(I-\Lambda)^{-1}$, respectively, corresponds to the concatenation of two directed paths at a common top node or by joining them with a bidirected edge. A top node represents a diagonal entry of $\Omega$, and a bidirected edge an off-diagonal entry of $\Omega$. Q.E.D.

**Example 4.3.** In Example 2.4, the coordinate function $(\phi_G)_{24}$ is a polynomial with four terms; see (2.8). The terms correspond to the four treks shown in Figure 4.1.

**Example 4.4.** Let $G$ be the graph from Figure 4.2, which contains the directed cycle $2 \to 3 \to 4 \to 2$. Due to this cycle, $\det(I-\Lambda) = 1 - \lambda_{23}\lambda_{34}\lambda_{42}$. As an example of a coordinate of $\phi_G$ we select

$$\begin{aligned}
&\phi_G(\Lambda,\Omega)_{24} = \\
&\quad \frac{1}{(1-\lambda_{23}\lambda_{34}\lambda_{42})^2}\left[\lambda_{12}^2\lambda_{23}\lambda_{34}\omega_{11} + \lambda_{12}\lambda_{13}\lambda_{34}\omega_{11}(\lambda_{23}\lambda_{34}\lambda_{42}+1)\right. \\
&\quad \left. + \lambda_{13}^2\lambda_{34}^2\lambda_{42}\omega_{11} + \lambda_{23}\lambda_{34}\omega_{22} + \lambda_{34}^2\lambda_{42}\omega_{33} + 2\lambda_{34}\lambda_{42}\omega_{34} + \lambda_{42}\omega_{44}\right].
\end{aligned}$$

To understand how this rational formula relates to the trek rule, let us focus on the treks from 2 to 4 that use the bidirected edge $2 \leftrightarrow 4$. There are then two treks for which both left and right side are self-avoiding

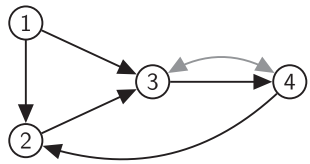


Fig. 4.2. A cyclic mixed graph.

paths, namely,

$$\tau_1 : 2 \leftarrow 4 \leftrightarrow 3 \rightarrow 4, \qquad\qquad \tau_2 : 2 \leftarrow 4 \leftarrow 3 \leftrightarrow 4.$$

Both of these treks yield the same monomial and together contribute the term $2\lambda_{34}\lambda_{42}\omega_{34}$ to $(\phi_G)_{24}$. All other treks from 2 to 4 that use edge $2 \leftrightarrow 4$ are obtained by inserting directed cycles into $\tau_1$ or $\tau_2$. For instance, inserting one cycle on the left- and one on the right-hand side of $\tau_1$ gives

$$2 \leftarrow 4 \leftarrow 3 \leftarrow 2 \leftarrow 4 \leftrightarrow 3 \rightarrow 4 \rightarrow 2 \rightarrow 3 \rightarrow 4.$$

The monomials associated with these treks are

$$\lambda_{34}\lambda_{42}\omega_{34}\left(\lambda_{23}\lambda_{34}\lambda_{42}\right)^k, \qquad k = 1, 2, \ldots .$$

The monomial for exponent $k$ arises from $k+1$ different treks; $l = 0, 1, \ldots, k$ cycles are inserted on the left, the remaining $k - l$ cycles are inserted on the right. Hence, the contribution to $(\phi_G)_{24}$ made by all treks from 2 to 4 that use edge $2 \leftrightarrow 4$ is

$$2\sum_{k=0}^{\infty}(k+1)\lambda_{34}\lambda_{42}\omega_{34}\left(\lambda_{23}\lambda_{34}\lambda_{42}\right)^k = \frac{2\lambda_{34}\lambda_{42}\omega_{34}}{(1-\lambda_{23}\lambda_{34}\lambda_{42})^2},$$

assuming that $|\lambda_{23}\lambda_{34}\lambda_{42}| < 1$. This explains one of the terms in the rational expression for $(\phi_G)_{24}$. The reasoning for the other terms is analogous.

## §5. Induced Subgraphs and Principal Submatrices

Suppose $X = (X_i : i \in V)$ follows the linear structural equation model given by mixed graph $G = (V, D, B)$, so $\mathrm{Var}[X] = \phi_G(\Lambda, \Omega)$ for some $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega \in PD(B)$. Let $A \subseteq V$ be a subset of nodes. Then the covariance matrix of the subvector $X_A = (X_i : i \in A)$ is obtained by projecting to the relevant principal submatrix, that is,

$$\mathrm{Var}[X_A] \;=\; \phi_G(\Lambda, \Omega)_{A,A}. \tag{5.1}$$

The resulting map $(\Lambda, \Omega) \mapsto \phi_G(\Lambda, \Omega)_{A,A}$ may be complicated, even when $\phi_G$ is not.

**Example 5.1.** Suppose $G = (V, D, \emptyset)$ is a directed graph with $i \to j \in D$ if and only if $i \notin A$ and $j \in A$; the graph is thus bipartite. Then the image of $\phi_G(\Lambda, \Omega)_{A,A}$ is the set of covariance matrices of a factor analysis model with $|V \setminus A|$ factors. Factor analysis models have complicated geometric structure, particularly when $V \setminus A$ has more than two elements [8, 20, 23, 62]. Open problems remain even for $|V \setminus A| = 1$ when allowing additional directed edges among the nodes in $A$ [47].

For a general mixed graph $G = (V, D, B)$, let $D_A = D \cap (A \times A)$ be the set of directed edges with both endpoints in $A$. Similarly, let $B_A \subset B$ be the set of bidirected edges that have both endpoints in $A$. The subgraph induced by $A$ is the mixed graph $G_A = (A, D_A, B_A)$. Example 5.1 and also already Example 1.1 show that the covariance matrices obtained from $\phi_{G_A}$ generally differ from those obtained by projecting $\phi_G$ onto the $A \times A$ submatrix. However, as we now emphasize, induced subgraphs are relevant in a special case.

Define the set of *parents* of a node $i \in V$ as

$$\mathrm{pa}(i) \;=\; \{j \in V \;:\; j \to i\}.$$

A set $A \subseteq V$ is *ancestral* if $i \in A$ implies $\mathrm{pa}(i) \subseteq A$. The terminology indicates that such a set contains all its ancestors, where an ancestor is a node from which there is a directed path to some node in $A$. Ancestral sets are obtained by recursively removing sink nodes from $V$. A *sink* is a node that is a head on all edges it is incident to. By our convention, both endpoints of bidirected edges are heads. Hence, node $i$ is a sink of $G = (V, D, B)$ if and only if $i$ is a sink of the directed part $(V, D)$.

**Theorem 5.2.** *Let $G = (V, D, B)$ be a mixed graph, and let $G_A$ be the subgraph induced by an ancestral set $A \subset V$. Then for all $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega \in PD(B)$, we have*

$$\phi_{G_A}(\Lambda_{A,A}, \Omega_{A,A}) \;=\; [\phi_G(\Lambda, \Omega)]_{A,A}\,.$$

*Proof.* Let $i, j \in A$. By the trek-rule, the $(i, j)$ entry of $\phi_G(\Lambda, \Omega)$ is given by summing the monomials associated to treks from $i$ to $j$ in $G$. Because $A$ is ancestral, a trek from $i$ to $j$ in $G$ cannot leave $A$. Hence, $G$

and $G_A$ have the same sets of treks from $i$ to $j$. Applying the trek-rule to $G_A$ yields the claim. Q.E.D.

**Example 5.3.** Take up Example 2.1. Node 3 is a sink in the graph from Figure 1.2 and the set $\{1, 2\}$ is ancestral. Inspecting the matrix displayed in (2.5), we see that removing the third row and column yields the matrix for the induced subgraph $1 \to 2$.

## §6. Graph Decomposition

We now present an important decomposition, which allows one to address several questions of interest by treating smaller subgraphs. The decomposition for acyclic mixed graphs was introduced by Jin Tian [65, 66]. Here we give the natural extension to graphs with directed cycles.

We consider two partitions of the vertex set of a mixed graph $G = (V, D, B)$. The first is given by the connected components of the bidirected part $(V, B)$. The second is given by the strongly connected components of the digraph $(V, D)$. Two distinct nodes belong to the same strongly connected component if there are directed paths in either direction between them. Let $\mathcal{C}(G)$ be the finest common coarsening of the two partitions. Two nodes are in the same block of $\mathcal{C}(G)$ if and only if they are connected by a path that uses only edges that are bidirected or part of some directed cycle. Note that $G$ is acyclic if and only if all strongly connected components are singleton sets. The blocks of $\mathcal{C}(G)$ are then simply the connected components of the bidirected part $(V, B)$.

For a block $C \in \mathcal{C}(G)$, define

$$V[C] := C \cup \bigcup_{i \in C} \mathrm{pa}(i)$$

to be the union of the block and all parents of nodes in the block. Let $D[C] = D \cap (V[C] \times C)$ be the set of directed edges with head in $C$. Let $B[C]$ be the set of bidirected with both endpoints in $C$.

**Definition 6.1.** *The graphs* $G[C] = (V[C], D[C], B[C])$, $C \in \mathcal{C}(G)$, *form a decomposition of* $G$, *and we refer to them as the* mixed components *of* $G$.

Graph decompositions partition edge sets. As we are working with mixed graphs both edge sets are partitioned in the decomposition. We note that the set $V[C] \setminus C$ contains the sources of $G[C]$. A source node is a tail on all edges it is incident to, with the convention that both endpoints of a bidirected edge are heads.

**Example 6.2.** The graph in Figure 6.1 has bidirected components $\{1,4\}$, $\{3\}$, and $\{2,5\}$. The strongly connected directed components are $\{1\}$, $\{2,3\}$, $\{4\}$, and $\{5\}$. The finest common coarsening of the two partitions is $\{\{1,4\},\{2,3,5\}\}$. The graph thus has two mixed components with vertex sets $V[\{2,3,5\}] = \{1,2,3,4,5\}$ and $V[\{1,4\}] = \{1,2,3,4\}$. The mixed component with vertex set $V[\{2,3,5\}]$ contains all edges with a head in $\{2,3,5\}$, and the second mixed component contains all edges with a head in $\{1,4\}$. The components are depicted in Figure 6.1.

For $C \in \mathcal{C}(G)$, define the projection $\pi_C^{\rightarrow} : \mathbb{R}^{V\times V} \to \mathbb{R}^{V[C]\times V[C]}$ by

$$\pi_C^{\rightarrow}(\Lambda)_{ij} = \begin{cases} \lambda_{ij} & \text{if } j \in C, \\ 0 & \text{if } j \in V[C] \setminus C. \end{cases}$$

Define a second map $\pi_C^{\leftrightarrow} : \mathbb{R}^{V\times V} \to \mathbb{R}^{V[C]\times V[C]}$ by

$$\pi_C^{\leftrightarrow}(\Omega)_{ij} = \begin{cases} \omega_{ij} & \text{if } i,j \in C, \\ 1 & \text{if } i = j \in V[C] \setminus C, \\ 0 & \text{otherwise.} \end{cases}$$

So $\pi_C^{\leftrightarrow}$ projects to the $C \times C$ submatrix and then adds the identity as diagonal block over $V[C] \setminus C$. Let $PD(B[C]) \subset \mathbb{R}^{V[C]\times V[C]}$ be the set of positive definite matrices supported over $B[C]$ and define the subset

$$PD_I(B[C]) \;=\; \{\Omega = (\omega_{ij}) \in PD(B[C]) \;:\; \omega_{ii} = 1 \;\text{ if }\; i \in V[C] \setminus C\}\,.$$

Then we have

$$\pi_C^{\rightarrow} : \mathbb{R}^D \to \mathbb{R}^{D[C]} \quad \text{and} \quad \pi_C^{\leftrightarrow} : PD(B) \to PD_I(B[C])$$

because $G[C]$ is a subgraph of $G$. With $\pi_C = (\pi_C^{\rightarrow}, \pi_C^{\leftrightarrow})$, we obtain the isomorphism

$$\pi = (\pi_C)_{C\in\mathcal{C}(G)} \;:\; \mathbb{R}^D \times PD(B) \;\to\; \prod_{C\in\mathcal{C}(G)} \mathbb{R}^{D[C]} \times PD_I(B[C]).$$

**Example 6.3.** Let $G$ be the graph from Figure 6.1, which has $\mathcal{C}(G) = \{\{1,4\},\{2,3,5\}\}$. A matrix in $\mathbb{R}^D$ is of the form

$$\Lambda = \begin{bmatrix} 0 & \lambda_{12} & \lambda_{13} & 0 & 0 \\ 0 & 0 & \lambda_{23} & \lambda_{24} & \lambda_{25} \\ 0 & \lambda_{32} & 0 & \lambda_{34} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

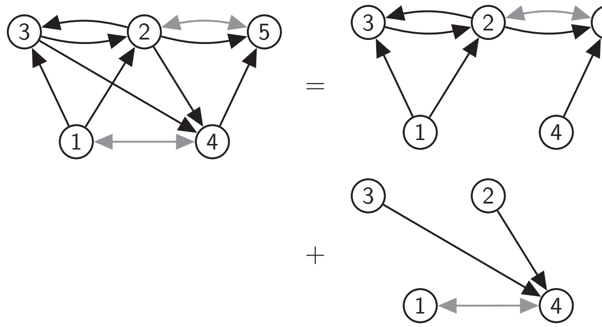


Fig. 6.1. A mixed graph is decomposed into its two mixed components.

Its projections are

$$\pi^{\rightarrow}_{\{2,3,5\}}(\Lambda) = \begin{bmatrix} 0 & \lambda_{12} & \lambda_{13} & 0 & 0 \\ 0 & 0 & \lambda_{23} & 0 & \lambda_{25} \\ 0 & \lambda_{32} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \pi^{\rightarrow}_{\{1,4\}}(\Lambda) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{24} \\ 0 & 0 & 0 & \lambda_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

A (symmetric) error covariance matrix in $PD(B)$ has the form

$$\Omega = \begin{bmatrix} \omega_{11} & 0 & 0 & \omega_{14} & 0 \\ 0 & \omega_{22} & 0 & 0 & \omega_{25} \\ 0 & 0 & \omega_{33} & 0 & 0 \\ \omega_{14} & 0 & 0 & \omega_{44} & 0 \\ 0 & \omega_{25} & 0 & 0 & \omega_{55} \end{bmatrix},$$

and we have

$$\pi^{\leftrightarrow}_{\{2,3,5\}}(\Omega) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & 0 & \omega_{25} \\ 0 & 0 & \omega_{33} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & \omega_{25} & 0 & 0 & \omega_{55} \end{bmatrix},$$

$$\pi^{\leftrightarrow}_{\{1,4\}}(\Omega) = \begin{bmatrix} \omega_{11} & 0 & 0 & \omega_{14} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \omega_{14} & 0 & 0 & \omega_{44} \end{bmatrix}.$$

With this preparation, Tian's theorem may be stated as follows. Recall that $PD_V$ denotes the cone of positive definite $V \times V$ matrices.

**Theorem 6.4.** *Let $G = (V, D, B)$ be a mixed graph with mixed components $G[C] = (V[C], D[C], B[C])$ for $C \in \mathcal{C}(G)$. Then there is an invertible map $\tau$ such that the following diagram commutes*:

$$\begin{tikzcd}
\mathbb{R}^{D}_{\mathrm{reg}} \times PD(B) \arrow[r, "\phi_G"] \arrow[d, shift left, "\pi"] & PD_V \arrow[d, shift left, "\tau"] \\
\prod_{C\in\mathcal{C}(G)} \mathbb{R}^{D[C]}_{\mathrm{reg}} \times PD_I(B[C]) \arrow[r, "(\phi_{G[C]})_{C\in\mathcal{C}(G)}"] \arrow[u, shift left] & \prod_{C\in\mathcal{C}(G)} PD_{V[C]} \arrow[u, shift left]
\end{tikzcd}$$

*In other words, $\tau \circ \phi_G = (\phi_{G[C]} \circ \pi_C)_{C\in\mathcal{C}(G)}$. Both $\tau$ and its inverse are rational maps, defined on all of $PD_V$ and all of $\prod_{k=1}^{m} PD_{V_k}$, respectively.*

Below we give a linear algebraic proof that makes $\tau$ and its rational nature explicit. Alternatively, a proof in probabilistic notation could be given by generalizing the proof of [66, Lemma 1]. The strongly connected components of $(V, D)$ would play the role of nodes in the setup of [66].

Theorem 6.4 is a very useful result as questions about $\phi_G$ can be answered by studying, one by one, the maps $\phi_{G[C]}$ for the mixed components. The fact that $\tau$ and $\tau^{-1}$ are rational is important. For instance, it allows one to obtain precise algebraic information about parameter identifiability in the sense of Definition 3.5.

**Corollary 6.5.** *The degree of identifiability of a mixed graph $G$ is the product of the degrees of identifiability of its mixed components $G[C]$, $C \in \mathcal{C}(G)$. In particular, $\phi_G$ is (generically) injective if and only if each $\phi_{G[C]}$ is so, for $C \in \mathcal{C}(G)$.*

Our proof of Theorem 6.4 is presented in terms of Cholesky decompositions. When applied to a Gaussian covariance matrix, the Cholesky decomposition corresponds to factoring the multivariate normal density into a product of conditional densities, which is the connection to the probabilistic setting of [66]. We begin by stating a lemma on uniqueness and sparsity in block-Cholesky decomposition.

If $\mathcal{C}$ is a partition of a finite set $V$, then we write $\mathrm{Diag}(\mathcal{C})$ for the space of matrices that are block-diagonal with respect to $\mathcal{C}$. So, $A = (a_{ij}) \in \mathbb{R}^{V\times V}$ is in $\mathrm{Diag}(\mathcal{C})$ if and only if $a_{ij} = 0$ whenever $i$ and $j$ are in distinct blocks of $\mathcal{C}$. If we order the blocks of the partition as $\mathcal{C} = \{C_1, \dots, C_k\}$, then we may define a space of strictly block upper-triangular matrices $\mathrm{Upper}(C)$, which contains $A = (a_{ij}) \in \mathbb{R}^{V\times V}$ if and only if $a_{ij} = 0$ whenever $i \in C_u$ and $j \in C_v$ with $u \geq v$.

**Lemma 6.6.** *Let $\Sigma \in PD_V$, and let $\mathcal{C}$ be a partition of $V$, with ordered blocks.*

(i) *There exist unique matrices $A \in \mathrm{Upper}(\mathcal{C})$ and $\Delta \in \mathrm{Diag}(\mathcal{C})$ such that*

$$\Sigma = (I - A)^{-T} \Delta (I - A)^{-1}.$$

*The matrix $\Delta$ has positive definite diagonal blocks.*

(ii) *Let $\mathcal{C}'$ be a second partition of $V$ that is coarser than $\mathcal{C}$. If $\Sigma \in \mathrm{Diag}(\mathcal{C}')$, then the matrix $A$ from* (i) *satisfies $A \in \mathrm{Diag}(\mathcal{C}')$.*

*Proof.* (i) A block-LDL decomposition yields $\Sigma = (I + L)\Delta(I + L)^T$ for unique $L^T \in \mathrm{Upper}(\mathcal{C})$ and $\Delta \in \mathrm{Diag}(\mathcal{C})$. Unit block upper-triangular matrices form a group and, thus, $(I + L)^{-T} = I - A$ for $A \in \mathrm{Upper}(\mathcal{C})$. (ii) The claim is a consequence of the way fill-in occurs in the Cholesky decomposition of a sparse matrix [68, Section 4.1]. Q.E.D.

*Proof of Theorem 6.4.* We first show that, for every block $C \in \mathcal{C}(G)$, there exists a map $\tau_C$ such that $\tau_C \circ \phi_G = \phi_{G[C]} \circ \pi_C$. Let $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$, $\Omega \in PD(B)$, and $\Sigma = \phi_G(\Lambda, \Omega)$. Then the claim is that

(6.1)

$$\tau_C(\Sigma) = \left[ I - \begin{pmatrix} 0 & \Lambda_{\overline{C},C} \\ 0 & \Lambda_{C,C} \end{pmatrix} \right]^{-T} \begin{pmatrix} I & 0 \\ 0 & \Omega_{C,C} \end{pmatrix} \left[ I - \begin{pmatrix} 0 & \Lambda_{\overline{C},C} \\ 0 & \Lambda_{C,C} \end{pmatrix} \right]^{-1}.$$

Let $\mathcal{C}(D)$ be the partition of $V$ given by the strongly connected components of $(V, D)$. Order the blocks of $\mathcal{C}(D)$ topologically as $W_1, \dots, W_k$ such that the existence of a directed path from a node in $W_u$ to a node in $W_v$ implies that $v \geq u$. By Lemma 6.6(i), there are $A \in \mathrm{Upper}(\mathcal{C}(D))$ and $\Delta \in \mathrm{Diag}(\mathcal{C}(D))$ such that

$$\Sigma = (I - A)^{-T} \Delta (I - A)^{-1}. \tag{6.2}$$

Letting $\overline{C} = V[C] \setminus C$, define

(6.3)

$$\tau_C(\Sigma) = \left[ I - \begin{pmatrix} 0 & A_{\overline{C},C} \\ 0 & A_{C,C} \end{pmatrix} \right]^{-T} \begin{pmatrix} I & 0 \\ 0 & \Delta_{C,C} \end{pmatrix} \left[ I - \begin{pmatrix} 0 & A_{\overline{C},C} \\ 0 & A_{C,C} \end{pmatrix} \right]^{-1}.$$

If $G$ is acyclic then $\Lambda$ is strictly upper-triangular under a topological ordering and, thus, $\Lambda \in \mathrm{Upper}(\mathcal{C}(D))$. When $G$ has directed cycles, then $\Lambda$ is block upper-triangular but not strictly so. Hence, we consider the block-diagonal matrix

$$\Delta_\Lambda \;=\; \mathrm{diag}(I - \Lambda_{W,W} : W \in \mathcal{C}(D)),$$

which is invertible because $\det(I-\Lambda) = \det(\Delta_\Lambda)$ and $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$. Hence,

$$\Sigma = \left[(I-\Lambda)\Delta_\Lambda^{-1}\right]^{-T}\left[\Delta_\Lambda^{-1}\Omega\Delta_\Lambda^{-1}\right]\left[(I-\Lambda)\Delta_\Lambda^{-1}\right]^{-1}. \tag{6.4}$$

Because $\Delta_\Lambda, \Omega \in \mathrm{Diag}(\mathcal{C}(G))$, we have

$$\tilde{\Omega} \;=\; \Delta_\Lambda^{-T}\Omega\Delta_\Lambda^{-1} \;\in\; \mathrm{Diag}(\mathcal{C}(G)).$$

Moreover, due to the block upper-triangular shape of $\Lambda$,

$$\tilde{\Lambda} \;=\; I-(I-\Lambda)\Delta_\Lambda^{-1} \;\in\; \mathrm{Upper}(\mathcal{C}(D)).$$

By Lemma 6.6(i) and (ii), there are matrices $\Delta_\Omega \in \mathrm{Diag}(\mathcal{C}(D))$ and $U \in \mathrm{Upper}(\mathcal{C}(D)) \cap \mathrm{Diag}(\mathcal{C}(G))$ such that

$$\tilde{\Omega} = (I-U)^{-T}\Delta_\Omega(I-U)^{-1}. \tag{6.5}$$

Combining (6.4) and (6.5) gives

$$\Sigma = \left[(I-\tilde{\Lambda})(I-U)\right]^{-T}\Delta_\Omega\left[(I-\tilde{\Lambda})(I-U)\right]^{-1}, \tag{6.6}$$

where $(I-\tilde{\Lambda})(I-U) = I-(\tilde{\Lambda}+U-\tilde{\Lambda}U)$ with $\tilde{\Lambda}+U-\tilde{\Lambda}U \in \mathrm{Upper}(\mathcal{C}(D))$.

By the uniqueness in Lemma 6.6(i), equations (6.2) and (6.6) imply

$$\Delta = \Delta_\Omega, \tag{6.7}$$

$$A = \tilde{\Lambda} + U - \tilde{\Lambda}U. \tag{6.8}$$

Since $U \in \mathrm{Diag}(\mathcal{C}(G))$, we have that

$$U_{V\times C} = \begin{pmatrix} U_{(V\setminus C)\times C} \\ U_{C\times C} \end{pmatrix} = \begin{pmatrix} 0 \\ U_{C\times C} \end{pmatrix}. \tag{6.9}$$

Therefore, by (6.8), $A_{C,C} = \tilde{\Lambda}_{C,C} + U_{C,C} - \tilde{\Lambda}_{C,C}U_{C,C}$. We deduce that

$$I - A_{C,C} = (I-\tilde{\Lambda}_{C,C})(I-U_{C,C}) \;\text{ and }\; A_{\overline{C},C} = \tilde{\Lambda}_{\overline{C},C}(I-U_{C,C}). \tag{6.10}$$

Moreover,

$$\tilde{\Omega}_{C,C} = (I-U_{C,C})^{-T}\Delta_{C,C}(I-U_{C,C})^{-1}, \tag{6.11}$$

which follows from (6.7) and the fact that

$$\left[(I-U)^{-1}\right]_{V,C} = \begin{pmatrix} 0 \\ \left[(I-U)^{-1}\right]_{C,C} \end{pmatrix} = \begin{pmatrix} 0 \\ (I-U_{C,C})^{-1} \end{pmatrix},$$

which in turn follows from $U$ being in $\mathrm{Diag}(\mathcal{C}(G))$.

Substituting the formulas from (6.10) into (6.3) yields that

$$\begin{aligned}(6.12)\quad \tau_C(\Sigma) = &\left[\begin{pmatrix} I & -\tilde{\Lambda}_{\overline{C},C} \\ 0 & I-\tilde{\Lambda}_{C,C}\end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & I-U_{C,C}\end{pmatrix}\right]^{-T} \\ &\times \begin{pmatrix} I & 0 \\ 0 & \Delta_{C,C}\end{pmatrix}\left[\begin{pmatrix} I & -\tilde{\Lambda}_{\overline{C},C} \\ 0 & I-\tilde{\Lambda}_{C,C}\end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & I-U_{C,C}\end{pmatrix}\right]^{-1}.\end{aligned}$$

Using (6.11), we get that
(6.13)

$$\begin{pmatrix} I & 0 \\ 0 & I-U_{C,C}\end{pmatrix}^{-T}\begin{pmatrix} I & 0 \\ 0 & \Delta_{C,C}\end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & I-U_{C,C}\end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ 0 & \tilde{\Omega}_{C,C}\end{pmatrix}.$$

Recalling the definition of $\Delta_\Lambda$ and $\tilde{\Lambda}$, we have

$$(6.14)\qquad \begin{pmatrix} I & -\tilde{\Lambda}_{\overline{C},C} \\ 0 & I-\tilde{\Lambda}_{C,C}\end{pmatrix} = \begin{pmatrix} I & -\Lambda_{\overline{C},C} \\ 0 & I-\Lambda_{C,C}\end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & (\Delta_\Lambda^{-1})_{C,C}\end{pmatrix}$$

Plugging (6.13) and (6.14) into (6.12), we obtain the claim from (6.1).

The entries of the matrices $A$ and $\Delta$ in (6.2) are rational functions of $\Sigma$ that are defined on all of $PD_V$. Hence, the same is true for the map $\tau_C$ defined in (6.3).

The value of $\tau_C(\Sigma)$ uniquely determines the matrices $A_{C,C}$, $A_{\overline{C},C}$, and $\Delta_{C,C}$ in (6.3). They are determined through a block LDL decomposition and, thus, rational functions of $\tau_C(\Sigma)$. Knowing the three matrices for all $C \in \mathcal{C}(G)$, we can form $A$ and $\Delta$ and recover $\Sigma$ using (6.2). We conclude that $\tau$ is invertible and the inverse is rational. Q.E.D.

# Part III. Parameter Identification

## §7. Global Identifiability

This section treats Question 3.1, which asks for a characterization of the mixed graphs $G = (V, D, B)$ for which the map $\phi_G$ is injective. In the statistical literature a model with injective parametrization is also called *globally identifiable*.

**Example 7.1.** If $G$ is the graph from Figure 3.1, then $\phi_G$ is injective. Indeed, the coefficients for the two directed edges of $G$ satisfy $\sigma_{11}\lambda_{12} = \sigma_{12}$ and $\sigma_{44}\lambda_{43} = \sigma_{34}$; recall (3.3). Since every positive definite matrix $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{4\times 4}$ has $\sigma_{11}, \sigma_{44} > 0$, these two equations always have a unique solution. Hence, all fibers $\mathcal{F}_G(\Lambda, \Omega)$ are singleton sets. In contrast, if $G$ is the graph from Figure 1.2, then only generic fibers are singleton sets and $\phi_G$ is not injective; recall Example 3.2.

Our first observation ties in with classical linear algebra.

**Theorem 7.2.** *If $G = (V, D, \emptyset)$ is an acyclic digraph, then $\phi_G$ is injective and has a rational inverse.*

We give two proofs. The first one emphasizes the connection to Cholesky decomposition.

*Proof A.* Suppose $V = \{1, \ldots, m\}$ is enumerated in reversed topological order such that $i \to j \in D$ implies that $j < i$. Then $\Lambda$ is a strictly lower-triangular matrix, and $\phi_G(\Lambda, \Omega)$ has matrix inverse $(I - \Lambda)\Omega^{-1}(I - \Lambda)^T$. This is the product of a unit lower-triangular matrix, a positive diagonal matrix and a unit upper-triangular matrix. We may compute $I - \Lambda$ and $\Omega^{-1}$ by an LDL decomposition. Q.E.D.

The second proof emphasizes the graphical nature of the problem and possible sparsity of $\Lambda$. It shows more explicitly that the inverse of $\phi_G$ is rational.

*Proof B.* Letting $\Sigma = (\sigma_{ij}) = \phi_G(\Lambda, \Omega)$, we have that

$$\Sigma_{\mathrm{pa}(i),i} \;=\; \Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}\Lambda_{\mathrm{pa}(i),i} \tag{7.1}$$

because if $j \in \mathrm{pa}(i)$, then every trek from $j$ to $i$ ends with an edge $k \to i$ for $k \in \mathrm{pa}(i)$. Indeed, a trek from $j$ to $i$ for which this fails has to be a directed path from $i$ to $j$. Adding the edge $j \to i$ to this path would yield a directed cycle. Similarly, every nontrivial trek from $i$ to $i$ begins and ends with a directed edge whose tail is a parent of $i$. Hence,

$$\sigma_{ii} \;=\; \omega_{ii} + \Lambda^T_{\mathrm{pa}(i),i}\Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}\Lambda_{\mathrm{pa}(i),i}. \tag{7.2}$$

The matrix $\Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}$ is a principal submatrix of the positive definite matrix $\Sigma$ and, thus, invertible. Therefore,

$$\Lambda_{\mathrm{pa}(i),i} \;=\; \left(\Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}\right)^{-1}\Sigma_{\mathrm{pa}(i),i}, \tag{7.3}$$

$$\omega_{ii} \;=\; \sigma_{ii} - \Sigma_{i,\mathrm{pa}(i)}\left(\Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}\right)^{-1}\Sigma_{\mathrm{pa}(i),i}. \tag{7.4}$$

Q.E.D.

Proof B shows that the formula from (7.3) holds more generally. It merely needs to hold that every trek from a node $j \in \mathrm{pa}(i)$ to $i$ ends with a directed edge that has $i$ as its head, so an edge of the form $k \to i$. This holds for every node in the graph if and only if the graph is *ancestral* [50]. A mixed graph is ancestral if the presence of a directed path from node $i$ to node $j$ implies that $i \neq j$ and $i \leftrightarrow j \notin B$. An ancestral graph is in particular acyclic.

The next easy lemma is crucial for the understanding of injectivity of $\phi_G$.

**Lemma 7.3.** *If $\phi_G$ is injective and $H \subseteq G$ is a* subgraph, *then $\phi_H$ is injective.*

*Proof.* Any subgraph can be obtained by removing edges one at a time, and then removing isolated nodes. If $H$ is obtained from $G = (V, D, B)$ by removing the edge $i \to j$, then $\phi_H$ is the restriction of $\phi_G$ to the subset of matrices $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ that have $\lambda_{ij} = 0$. If we instead remove the edge $i \leftrightarrow j$, then the restriction is to matrices $\Omega \in PD(B)$ with $\omega_{ij} = 0$. If $H$ is obtained by removing the isolated node $i$, then

$$\phi_G(\Lambda, \Omega) = \begin{pmatrix} \phi_H(\Lambda, \Omega) & 0 \\ 0 & \omega_{ii} \end{pmatrix}.$$

In either case non-injectivity of $\phi_H$ implies non-injectivity of $\phi_G$. Q.E.D.

**Corollary 7.4.** *If $\phi_G$ is injective, then $G$ is simple, that is, for any two distinct vertices $i$ and $j$ at most one of the three edges $i \leftrightarrow j$, $i \to j$ and $i \leftarrow j$ may appear in $G$.*

*Proof.* If $G$ is not simple then it contains a subgraph $H$ with two nodes and two edges. The map $\phi_H$ is infinite-to-one as it maps a four-dimensional domain into the three-dimensional set of symmetric $2 \times 2$ matrices. Now apply Lemma 7.3. Q.E.D.

**Theorem 7.5.** *If $\phi_G$ is injective, then $G$ is acyclic.*

*Proof sketch.* By Lemma 7.3, we may restrict to studying directed cycles $1 \to 2 \to \dots \to m \to 1$. The case of $m = 2$ is covered by Corollary 7.4. If $m \geq 3$, then it is possible to show that $\phi_G$ is generically 2-to-1, that is, the fiber $\mathcal{F}_G(\Lambda, \Omega)$ is generically of size two [16]. Q.E.D.

It remains to characterize injectivity for acyclic graphs $G = (V, D, B)$. The next theorem shows that injectivity can be decided in polynomial time by alternatingly decomposing the bidirected part $(V, B)$ into connected components and removing sink nodes of the directed part $(V, D)$.

**Theorem 7.6.** *Suppose $G$ is an acyclic mixed graph. Then*:

(a) *$\phi_G$ is injective if and only if $G$ does not contain a subgraph whose bidirected part is connected and whose directed part has a unique sink.*

(b) *If $\phi_G$ is injective, then its inverse is rational and $\mathcal{M}_G$ smooth.*

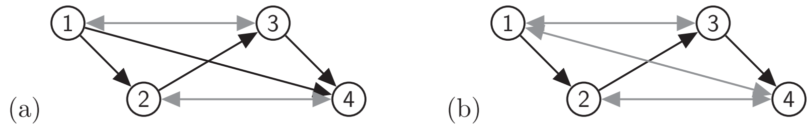


Fig. 7.1. (a) A mixed graph for which the parametrization $\phi_G$ is injective. (b) A graph for which $\phi_G$ is not injective.

Figure 7.1 illustrates the characterization in part (a) of the theorem. A full proof of the theorem can be found in [16]. The fact that $\phi_G$ is not injective if the combinatorial condition in (a) fails can be shown by a counterexample for the particular subgraph and then invoking Lemma 7.3. The sufficiency of the condition for injectivity can be proven by repeatedly applying the graph decomposition result in Theorem 6.4 and the result on ancestral subgraphs from Theorem 5.2. These results as well as Theorem 7.6 have generalizations to nonlinear structural equation models [56, 66].

## §8. Generic Identifiability

The difference between injectivity and generic injectivity of $\phi_G$ may appear minute. However, the two properties are quite different, and failure of generic injectivity cannot be argued by studying subgraphs (as in Lemma 7.3). According to Corollary 7.4, a mixed graph $G$ can have the map $\phi_G$ injective only if it is acyclic and simple. The deeper issue is then to find out which simple acyclic mixed graphs have $\phi_G$ injective. In contrast, the next result shows that all simple acyclic mixed graphs are generically injective. The deeper issue for generic injectivity is thus the treatment of graphs that contain directed cycles or are not simple.

**Theorem 8.1.** *If $G = (V, D, B)$ is acyclic and simple, then $\phi_G$ is generically injective and algebraically one-to-one.*

The theorem is due to [7]. It shows that the graph from Figure 7.1(b) has $\phi_G$ generically injective, but not injective. A short proof of Theorem 8.1 is obtained from the following observation.

**Lemma 8.2.** *Let $G = (V, D, B)$ be a mixed graph, and let $\Sigma = \phi_G(\Lambda_0, \Omega_0)$ for $\Lambda_0 \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega_0 \in PD(B)$. The fiber $\mathcal{F}_G(\Lambda_0, \Omega_0)$ is isomorphic to the set of matrices $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ that solve the equation system*

$$\big[(I-\Lambda)^T \Sigma (I-\Lambda)\big]_{ij} \;=\; 0, \qquad i \neq j,\ i \leftrightarrow j \notin B. \tag{8.1}$$

*Proof.* The projection $(\Lambda, \Omega) \mapsto \Lambda$ maps $\mathcal{F}_G(\Lambda_0, \Omega_0)$ to the set of matrices $\Lambda \in \mathbb{R}^D_{\text{reg}}$ that solve the equations in (8.1). Indeed, as $I - \Lambda$ is invertible for $\Lambda \in \mathbb{R}^D_{\text{reg}}$,

$$\Sigma = \phi_G(\Lambda, \Omega) = (I - \Lambda)^{-T}\Omega(I - \Lambda) \implies \Omega = (I - \Lambda)^T\Sigma(I - \Lambda).$$

If $i \neq j$ and $i \leftrightarrow j \notin B$, then the $(i, j)$ entry of $\Omega$ is zero. Conversely, if $\Lambda$ solves (8.1), then $(\Lambda, (I - \Lambda)^T\Sigma(I - \Lambda)) \in \mathcal{F}_G(\Lambda, \Omega)$. Q.E.D.

We emphasize that the equations in (8.1) are bilinear as

$$\begin{aligned} &\big[(I - \Lambda)^T\Sigma(I - \Lambda)\big]_{ij} \\ &\qquad = \sigma_{ij} - \sum_{k \in \text{pa}(i)} \lambda_{ki}\sigma_{ki} - \sum_{l \in \text{pa}(j)} \sigma_{il}\lambda_{lj} + \sum_{k \in \text{pa}(i)} \sum_{l \in \text{pa}(j)} \lambda_{ki}\sigma_{kl}\lambda_{lj}. \end{aligned}$$

*Proof of Theorem 8.1.* Because $G$ is acyclic, we may enumerate the vertex set in a topological order as $V = \{1, \ldots, m\}$. Then $\text{pa}(i) \subseteq \{1, \ldots, i-1\}$ for $i = 1, \ldots, m$. Moreover, because $G$ is simple, $j \in \text{pa}(i)$ implies that $j \leftrightarrow i \notin B$. By Lemma 8.2,

$$\big[(I - \Lambda)^T\Sigma(I - \Lambda)\big]_{\text{pa}(i),i} = 0, \qquad i = 1, \ldots, m.$$

These equations can be rewritten as

$$\big[(I - \Lambda)^T\Sigma\big]_{\text{pa}(i),\text{pa}(i)} \Lambda_{\text{pa}(i),i} = \big[(I - \Lambda)^T\Sigma\big]_{\text{pa}(i),i}, \quad i = 1, \ldots, m. \tag{8.2}$$

By the topological order, if $j \in \text{pa}(i)$, then the $j$-th row of $(I - \Lambda)^T\Sigma$ depends only on the first $i-1$ columns of $\Lambda$. The system in (8.2) can thus be solved recursively, each step requiring solution of a linear system.

To show that $\phi_G$ is generically injective, it remains to argue that the equations in (8.2) generically have a unique solution. It suffices to exhibit a single pair $(\Lambda, \Omega)$ for which this is true. We may choose $\Lambda = 0$ and $\Omega = I$, so $\Sigma = I$. Then the matrix for the $i$-th group of equations in (8.2) is $\Sigma_{\text{pa}(i),\text{pa}(i)}$, which is invertible. Q.E.D.

Although a combinatorial characterization of the graphs with generically injective parametrization $\phi_G$ is not known, Gröbner basis techniques can be used to determine the degree of identifiability from Definition 3.5 and, thus, decide whether $\phi_G$ is algebraically one-to-one. Gröbner bases are computationally tractable for non-trivial examples and have been used for a classification of all graphs with up to 5 nodes [34]. For larger graphs, algebraic methods can be applied after decomposition according to Theorem 6.4.

We describe two options for the computation. In either case, we advocate working with the equation system from (8.1) as opposed to the fiber equation $\Sigma = \phi_G(\Lambda, \Omega)$. System (8.1) has $\Omega$ eliminated and may be far more compact as it avoids inversion of $I - \Lambda$. This said, although system (8.1) is polynomial also for graphs that contain directed cycles, care must be taken to avoid spurious solutions with $I - \Lambda$ non-invertible.

The first possibility is to perform a parametric Gröbner basis computation. We introduce a matrix $\Lambda$ whose nonzero entries $\lambda_{ij}$, $i \to j \in D$, are indeterminates and a pair of matrices $(\Lambda_0, \Omega_0)$ that are parameters. We form the matrix $(I - \Lambda)^T \phi_G(\Lambda_0, \Omega_0)(I - \Lambda)$ and set to zero the off-diagonal entries indexed by non-edges of the bidirected part $(V, B)$. We then compute a Gröbner basis for the resulting system in the polynomial ring with coefficients in the field of rational fractions $\mathbb{R}(\Lambda_0, \Omega_0)$. The Gröbner basis readily yields the dimension of the generic fibers. If the dimension is finite we may also find the algebraic degree of the generic fibers, which is what we referred to as degree of identifiability. When the graph $G$ contains directed cycles, we first saturate our equation system with respect to $\det(I - \Lambda)$ in order to remove solutions that have $I - \Lambda$ non-invertible.

**Example 8.3.** The following code for the system Singular [13] implements the approach just described for a directed 3-cycle:

```
LIB "linalg.lib"; option(redSB);
ring R = (0,l012,l023,l031,w011,w022,w033),(l12,l23,l31),dp;
matrix L[3][3] = 1,-l12,0,
                0,1,-l23,
                -l31,0,1;
matrix L0[3][3] = 1,-l012,0,
                 0,1,-l023,
                 -l031,0,1;
matrix W0[3][3] = w011,0,0,
                 0,w022,0,
                 0,0,w033;
matrix W[3][3] =
  transpose(L)*inverse(transpose(L0))*W0*inverse(L0)*L;
ideal GB = sat(ideal(W[1,2],W[1,3],W[2,3]), det(L))[1];
dim(GB); mult(GB);
```

The output first certifies that the fibers are generically zero-dimensional, that is, contain finitely many points. The multiplicity computed with the last command shows the degree of identifiability to be two.

The second possibility is to consider only polynomials with real-valued coefficients but to introduce as polynomial variables the nonzero entries of $\Lambda$ as well as a symmetric matrix $\Sigma$. These variables are ordered with respect to a block monomial order in which the variables in $\Lambda$ are larger than the variables in $\Sigma$. Let $\mathcal{I}$ be the ideal generated by the off-diagonal entries of $(I-\Lambda)^T\Sigma(I-\Lambda)$ that are indexed by the non-edges of $(V,B)$. Saturate $\mathcal{I}$ with respect to $\det(I-\Lambda)$. Let $\mathcal{J}$ be the reduced Gröbner basis of the resulting ideal. Elimination theory yields the following fact [34, Section 8 of the supplemental material].

**Proposition 8.4.** *A mixed graph $G=(V,D,B)$ has $\phi_G$ algebraically one-to-one if and only if for each $i\to j\in D$, the reduced Gröbner basis $\mathcal{J}$ contains an element with leading monomial $a(\Sigma)\lambda_{ij}$.*

In comparison to the first approach, the second method yields relations that show how to identify coefficients $\lambda_{ij}$ from $\Sigma$. By analyzing the monomials under the staircase of the initial ideal of $\mathcal{J}$ [12, Chapter 9], we may also determine the generic dimension and degree of the fiber $\mathcal{F}_G(\Lambda,\Omega)$. This way we may find the degree of identifiability of $G$.

**Example 8.5.** Treating again a directed 3-cycle, we give an example of the second type of computation in Singular:

```
LIB "linalg.lib"; option(redSB);
ring R = 0,(l12,l23,l31,s11,s12,s13,s22,s23,s33),(dp(3));
matrix L[3][3] = 1,-l12,0,
                 0,1,-l23,
                 -l31,0,1;
matrix S[3][3] = s11,s12,s13,
                 s12,s22,s23,
                 s13,s23,s33;
matrix W[3][3] = transpose(L)*S*L;
ideal GB = sat(ideal(W[1,2],W[1,3],W[2,3]), det(L))[1]; GB;
```

The output is a list of 9 polynomials whose leading terms are, in our usual notation,

$$\begin{array}{ccc}
\lambda_{23}\lambda_{31}\sigma_{23}, & \lambda_{12}\lambda_{31}\sigma_{13}, & \lambda_{12}\lambda_{23}\sigma_{12}, \\
\lambda_{12}\lambda_{31}\sigma_{12}\sigma_{33}, & \lambda_{12}\sigma_{11}\sigma_{13}\sigma_{23}, & \lambda_{12}\lambda_{23}\sigma_{11}\sigma_{23}, \\
\lambda_{23}\sigma_{12}\sigma_{13}\sigma_{22}, & \lambda_{23}\lambda_{31}\sigma_{13}\sigma_{22}, & \lambda_{31}^2\sigma_{13}\sigma_{22}\sigma_{33}.
\end{array}$$

By Proposition 8.4, $\phi_G$ is not algebraically one-to-one because there is no leading term of the form $\lambda_{31}a(\Sigma)$. The last leading term belongs to a polynomial that shows that $\lambda_{31}$ is algebraic function of degree 2 of the covariance matrix $\Sigma$ because it solves the equation

$$\lambda_{31}^2\sigma_{33}(\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23})-\lambda_{31}(\sigma_{13}^2\sigma_{22}-\sigma_{11}\sigma_{23}^2-\sigma_{12}^2\sigma_{33}+\sigma_{11}\sigma_{22}\sigma_{33}) \\ +\sigma_{11}(\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23}) \;=\; 0.$$

The equations with leading terms $\lambda_{23}\sigma_{12}\sigma_{13}\sigma_{22}$ and $\lambda_{12}\sigma_{11}\sigma_{13}\sigma_{23}$ show that $\lambda_{23}$ and $\lambda_{12}$ are rational functions of $\Sigma$ and $\lambda_{31}$. Altogether, we have verified that $\phi_G$ is algebraically 2-to-one. Checking this by counting monomials under the staircase means considering the leading monomials while focusing only the variables $\lambda_{12}, \lambda_{23}, \lambda_{31}$ we seek to solve for. The monomials are

$$\lambda_{23}\lambda_{31}, \qquad \lambda_{12}\lambda_{31}, \qquad \lambda_{12}\lambda_{23}, \qquad \lambda_{12}, \qquad \lambda_{23}, \qquad \lambda_{31}^2. \tag{8.3}$$

They generate the ideal $\mathcal{I} = \langle \lambda_{12}, \lambda_{23}, \lambda_{31}^2 \rangle$. The monomials under the staircase are the monomials in $\mathbb{R}[\lambda_{12}, \lambda_{23}, \lambda_{31}] \setminus \mathcal{I}$. The fact that there are two, namely, 1 and $\lambda_{31}$, implies that $\phi_G$ is algebraically 2-to-one.

Although Gröbner basis methods can be effective, it is desirable to obtain combinatorial methods that are efficient also for large-scale problems. The half-trek criteria of [34] are state-of-the-art methods whose conditions can be checked in time that is polynomial in the size of the vertex set of the considered graph. They provide a sufficient as well as a necessary condition for generic injectivity of $\phi_G$. More precisely, there is a condition that is sufficient for $\phi_G$ to be algebraically one-to-one and a related condition that is necessary for $\phi_G$ to be generically finite-to-one. The conditions are implemented in a package for the R project for statistical computing [4]. We begin our discussion of the half-trek criteria by introducing some needed terminology.

A *half-trek* from initial node $i$ to target node $j$ is a trek $\tau$ from $i$ to $j$ whose left-hand side is a singleton set, so $\mathrm{left}\,(\tau) = \{i\}$. In other words, a half-trek is of the form

$$i \to j_1 \to \ldots \to j_r \to j \quad \text{or} \quad i \leftrightarrow j_1 \to \ldots \to j_r \to j.$$

Let $X, Y \subseteq V$ be two sets of nodes of equal cardinality $|X| = |Y| = k$. Let $\Pi$ be a set of $k$ treks. Then $\Pi$ is a system of treks from $X$ to $Y$, denoted $\Pi : X \rightrightarrows Y$, if $X$ is the set of initial nodes of the treks in $\Pi$ and $Y$ is the set of target nodes. Note that we allow $X \cap Y \neq \emptyset$. The system $\Pi$ is a system of half-treks if every trek $\pi_i$ is a half-trek. Finally, the system $\Pi$ has no sided intersection if

$$\mathrm{left}\,(\pi) \cap \mathrm{left}\,(\pi') = \emptyset = \mathrm{right}\,(\pi) \cap \mathrm{right}\,(\pi')$$

for all pairs of treks $\pi, \pi' \in \Pi$.

**Definition 8.6.** *A set $Y \subseteq V$ satisfies the* half-trek criterion *with respect to node $i$ if* (i) $|Y| = |\,\mathrm{pa}(i)|$, (ii) *$j = i$ or $j \leftrightarrow i$ implies that $j \notin Y$, and* (iii) *there exists a system of half-treks* $\Pi : Y \rightrightarrows \mathrm{pa}(i)$ *that has no sided intersection.*

**Theorem 8.7.** *Let $G = (V, D, B)$ be a mixed graph.*

(i) *Suppose for every $i \in V$ there exists a set $Y_i \subseteq V$ that satisfies the half-trek criterion with respect to $i$. If there exists a total ordering $\prec$ such that $j \prec i$ whenever $j \in Y_i$ and there is a half-trek from $i$ to $j$, then $\phi_G$ is generically injective and algebraically one-to-one.*

(ii) *For $\phi_G$ to be generically finite-to-one it is necessary that there exists a family of sets $Y_i \subseteq V$, $i \in V$, such that $Y_i$ satisfies the half-trek criterion with respect to $i$ and $j \in Y_i$ implies $i \notin Y_j$.*

We merely outline the proof of the theorem; for details see [34]. Some of the arguments are further illustrated in Example 8.8. Note also that Theorem 8.1 is obtained from Theorem 8.7(i) by taking $Y_i = \mathrm{pa}(i)$ and $\prec$ as a topological order.

*Outline of proof of Theorem 8.7.* (i) Let $\Sigma = \phi_G(\Lambda_0, \Omega_0)$ for $\Lambda_0 \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega_0 \in PD(B)$. Suppose $(\Lambda, \Omega) \in \mathcal{F}_G(\Lambda_0, \Omega_0)$. To show that $(\Lambda, \Omega) = (\Lambda_0, \Omega_0)$, we visit the nodes $i \in V$ from smallest to largest in the order $\prec$ and iteratively find a linear equation system that is uniquely solved by the $i$-th column of $\Lambda$. The starting point is Lemma 8.2, by which we have

$$\left[(I - \Lambda)^T \Sigma (I - \Lambda)\right]_{Y_i, i} = 0, \quad i \in V. \tag{8.4}$$

This is true because Definition 8.6 yields that $j \neq i$ and $j \leftrightarrow i \notin B$ when $j \in Y_i$. Similar to the proof of Theorem 8.1, we may rearrange (8.4) to

$$A_i(\Lambda, \Sigma)\Lambda_{\mathrm{pa}(i),i} \;=\; b_i(\Lambda, \Sigma),$$

with $A_i(\Lambda, \Sigma) = \left[(I - \Lambda)^T \Sigma\right]_{Y_i, \mathrm{pa}(i)}$ and $b_i(\Lambda, \Sigma) = \left[(I - \Lambda)^T \Sigma\right]_{Y_i, i}$. Both $A_i(\Lambda, \Sigma)$ and $b_i(\Lambda, \Sigma)$ can be shown to depend only on those columns of $\Lambda$ that are indexed by nodes $j$ with a half-trek from $i$ to $j$. Thus, the proof is complete if we can show that $A_i(\Lambda_0, \Sigma)$ is invertible for generic choices of $\Lambda_0$ and $\Omega_0$. To verify this, we may use the existence of a half-trek system without sided intersection from $Y_i$ to $\mathrm{pa}(i)$ to argue that the determinant of $A_i(\Lambda_0, \Sigma)$ is not the zero polynomial. This last step is in the spirit of the Lindström-Gessel-Viennot lemma.

(ii) The Jacobian of the equations from Lemma 8.2 has its rows indexed by the non-edges of the bidirected part $(V, B)$ and its columns indexed by the directed edges in $D$. For $\phi_G$ to be generically finite-to-one, it is necessary that the Jacobian has full column rank $D$. It can be shown that the Jacobian contains an invertible $|D| \times |D|$ submatrix only if the given condition holds. Let $J_i$ be the submatrix of the Jacobian

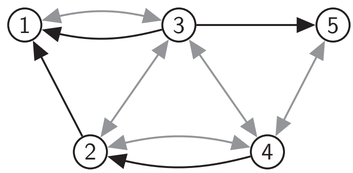


Fig. 8.1. Illustration of Theorem 8.7.

obtained by selecting the columns corresponding to directed edges with head at $i$. Then, more specifically, $J_i$ has full column rank only if there exists a subset $Y_i \subseteq V$ that satisfies the half-trek criterion with respect to $i$. Moreover, if $j \in Y_i$ and $i \in Y_j$, then the same row, namely, that corresponding to $i \leftrightarrow j \notin B$, would be used to get an invertible square submatrix of $J_i$ and $J_j$. Q.E.D.

The conditions from Theorem 8.7 can be checked in polynomial time. For condition (i), we may use recursively that the existence of a suitable set satisfying the half-trek criterion is equivalent to a condition on the maximum flow in a certain network-flow problem. Condition (ii) can be checked via a single larger network-flow problem [34, Section 6].

**Example 8.8.** Let $G$ be the graph in Figure 8.1. The sets

$$Y_1 = \{2, 5\}, \qquad Y_2 = \{5\}, \qquad Y_3 = \emptyset, \qquad Y_4 = \emptyset, \qquad Y_5 = \{3\}$$

each satisfy the half-trek criterion with respect to the node they are indexed by. This is least evident for $Y_1$, and we highlight the half-treks $2 \leftrightarrow 3$ and $5 \leftrightarrow 4 \to 2$, which have no sided intersection. Choosing the ordering as $3 \prec 4 \prec 5 \prec 1 \prec 2$, Theorem 8.7(i) shows that $\phi_G$ is algebraically one-to-one. Other possible orderings are obtained by permuting $\{3, 4, 5\}$ or $\{1, 2\}$.

To illustrate ideas from the proof of Theorem 8.7(i), we focus on node 1, with $\mathrm{pa}(1) = \{2, 3\}$. Since $Y_1 = \{2, 5\}$, we work with the equations

$$\big[(I-\Lambda)^T\Sigma(I-\Lambda)\big]_{51} \;=\; 0, \qquad \big[(I-\Lambda)^T\Sigma(I-\Lambda)\big]_{21} \;=\; 0.$$

Expanding out the matrix product, the equations become

$$\sigma_{51} - \sigma_{52}\lambda_{21} - \sigma_{53}\lambda_{31} - \lambda_{35}\sigma_{31} + \lambda_{35}\sigma_{32}\lambda_{21} + \lambda_{35}\sigma_{33}\lambda_{31} \;=\; 0, \tag{8.5}$$

$$\sigma_{21} - \sigma_{22}\lambda_{21} - \sigma_{23}\lambda_{31} - \lambda_{42}\sigma_{41} + \lambda_{42}\sigma_{42}\lambda_{21} + \lambda_{42}\sigma_{43}\lambda_{31} \;=\; 0, \tag{8.6}$$

and we wish to solve for $\lambda_{21}$ and $\lambda_{31}$. With $5 \prec 1$, we have already solved for $\lambda_{35}$; since $Y_5 = \mathrm{pa}(5)$ we have $\lambda_{35} = \sigma_{35}/\sigma_{55}$ as is also clear from the discussion after Proof B for Theorem 7.2. Substituting the ratio for $\lambda_{35}$ turns (8.5) into a linear equation in $\lambda_{21}$ and $\lambda_{31}$. The equation in (8.6) could be linearized similarly, except that now the relevant coefficient $\lambda_{42}$ has not yet been determined in an ordering with $1 \prec 2$. However, if $\Lambda$ is part of a pair $(\Lambda, \Omega)$ in the fiber given by $\Sigma$, then

$$-\lambda_{42}\sigma_{41} + \lambda_{42}\sigma_{42}\lambda_{21} + \lambda_{42}\sigma_{43}\lambda_{31} \;=\; 0$$

because there is no half-trek from 1 to 2. To see this note that the term $\lambda_{42}\sigma_{41}$ corresponds to treks from 2 to 1 that start with the edge $2 \leftarrow 4$, whereas the sum $\lambda_{42}\sigma_{42}\lambda_{21} + \lambda_{42}\sigma_{43}\lambda_{31}$ corresponds to treks from 2 to 1 that start with the edge $2 \leftarrow 4$ and end in either $2 \to 1$ or $3 \to 1$. These two sets of treks coincide when there is no half-trek from 1 to 2.

Condition (i) in Theorem 8.7 can be improved by applying the graph decomposition from Section 6 and checking the condition in each subgraph. No such strengthening is possible for the necessary condition from part (ii) of the theorem [34]. Further strengthening of the sufficient condition is possible by first removing sink nodes from the graph using the observation from Theorem 5.2. When a sink node is removed a more refined graph decomposition may become possible; we refer the reader to [10, 22]. While a specific polynomial-time algorithm using this idea is given in [22], it is still unclear how to best design algorithms based on recursive graph decomposition and removal of sink nodes.

We conclude our discussion of parameter identification with two examples from the exhaustive computational study of graphs with up to 5 nodes in [34]. Both graphs in Figure 8.1 satisfy the necessary condition in Theorem 8.7(ii) and, thus, have $\phi_G$ generically finite-to-one. Neither graph satisfies the sufficient condition from Theorem 8.7(i). The graph in panel (a) indeed does not have $\phi_G$ generically injective. Instead, $\phi_G$ is algebraically 3-to-one. The graph in panel (b), however, is algebraically one-to-one but Theorem 8.7(i) fails to recognize it. Decomposition and removal of sink nodes do not help and other ideas are needed; see [71].

# Part IV. Relations Among Covariances

## §9. Implicitization

Let $\mathcal{M}_G$ be the set of covariance matrices of the structural equation model given by a mixed graph $G = (V, D, B)$. Motivated, in particular, by the covariance equivalence problem from Question 3.7, we now discuss polynomial relations among the entries of the matrices in $\mathcal{M}_G$. Let

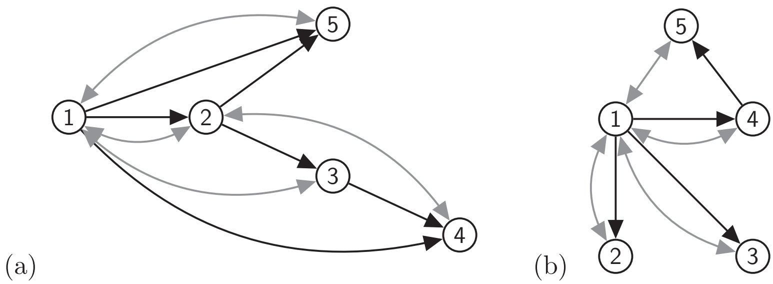


Fig. 8.2. Graphs that satisfy the necessary but not the sufficient condition from Theorem 8.7: (a) $\phi_G$ is algebraically 3-to-one, (b) $\phi_G$ is algebraically one-to-one and, thus, generically injective.

$\Sigma = (\sigma_{ij})$ be a symmetric $V \times V$ matrix of variables, and let $\mathbb{R}[\Sigma]$ be the ring of polynomials in the indeterminates $\sigma_{ij}$ with real coefficients. The relations we seek to understand make up the *vanishing ideal*

$$\mathcal{I}(G) \;=\; \{f \in \mathbb{R}[\Sigma] \;:\; f(\Sigma) = 0 \text{ for all } \Sigma \in \mathcal{M}_G\}\,.$$

Since $\mathcal{M}_G$ is the image of the rational map $\phi_G$, we may compute a generating set for $\mathcal{I}(G)$ by implicitization. Assume for simplicity that $G$ is acyclic and, thus, $\phi_G$ polynomial. Define $\mathbb{R}[\Sigma, \Lambda, \Omega]$ to be the polynomial ring with the additional indeterminates from a $V \times V$ matrix $\Lambda = (\lambda_{ij})$ and a symmetric $V \times V$ matrix $\Omega = (\omega_{ij})$ whose supports are determined by $D$ and $B$, respectively. Then $\langle \sigma_{ij} - \phi_G(\Lambda, \Omega)_{ij} : i \le j \rangle \subset \mathbb{R}[\Sigma, \Lambda, \Omega]$ is the ideal for the graph of $\phi_G$, and

$$\mathcal{I}(G) \;=\; \langle\; \sigma_{ij} - \phi_G(\Lambda, \Omega)_{ij} : i \le j \;\rangle \;\cap\; \mathbb{R}[\Sigma].$$

A better approach, however, is to start with the equations from Lemma 8.2, which have $\Omega$ already eliminated. We compute

$$\mathcal{I}(G) \;=\; \Big\langle \big[(I-\Lambda)^T \Sigma (I-\Lambda)\big]_{ij} \;:\; \; i \neq j,\; i \leftrightarrow j \notin B \Big\rangle \;\cap\; \mathbb{R}[\Sigma].$$

If $G$ is not acyclic, we saturate with respect to $\det(I - \Lambda)$ before intersecting with $\mathbb{R}[\Sigma]$; compare Examples 8.3 and 8.5.

**Example 9.1.** To illustrate the use of a different software, we change to Macaulay2 [40]. The following code computes a generating set for the vanishing ideal $\mathcal{I}(G)$ when $G$ is the graph from Figure 9.1:

```
R = QQ[l12,l13,l24,l34, s11,s12,s13,s14,s22,s23,s24,s33,s34,s44,
      MonomialOrder => Eliminate 4];
```

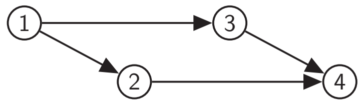

Fig. 9.1. An acyclic digraph.

```
Lambda = matrix{{1, -l12, -l13, 0},
                {0, 1, 0, -l24},
                {0, 0, 1, -l34},
                {0, 0, 0, 1}};
S = matrix{{s11, s12, s13, s14},
           {s12, s22, s23, s24},
           {s13, s23, s33, s34},
           {s14, s24, s34, s44}};
W = transpose(Lambda)*S*Lambda;
I = ideal{W_(0,1),W_(0,2),W_(0,3),W_(1,2),W_(1,3),W_(2,3)};
eliminate({l12,l13,l24,l34},I)
```

The ‘GraphicalModels’ package [37] automates the computation:

```
needsPackage "GraphicalModels";
G = digraph {{1,{2,3}},{2,{4}},{3,{4}}};
R = gaussianRing G;
gaussianVanishingIdeal R
```

Reproduced in our notation, the output shows that the ideal $\mathcal{I}(G)$ is generated by the two polynomials

$$f_1 = \sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23}, \tag{9.1}$$

$$\begin{aligned} f_2 = {} & \sigma_{14}\sigma_{23}^2 - \sigma_{13}\sigma_{23}\sigma_{24} - \sigma_{14}\sigma_{22}\sigma_{33} \\ & + \sigma_{12}\sigma_{24}\sigma_{33} + \sigma_{13}\sigma_{22}\sigma_{34} - \sigma_{12}\sigma_{23}\sigma_{34}. \end{aligned} \tag{9.2}$$

Computing $\mathcal{I}(G)$ using Gröbner bases is a method that applies to any mixed graph but can be computationally prohibitive for graphs with more than 5 or 6 nodes. In order to solve model equivalence problems combinatorial insight on particular types of relations is needed.

**Example 9.2.** Continuing with Example 9.1, observe that $f_1$ and $f_2$ from (9.1) and (9.2) are determinants of submatrices of the covariance matrix $\Sigma$. The following two displays locate the submatrices:

$$\begin{bmatrix} \boxed{\sigma_{11}} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix} \qquad \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}$$

The boxes emphasize that each submatrix contains a principal submatrix along with one additional row and column. The determinants $f_1$ and $f_2$ are thus almost principal minors of $\Sigma$. As we discuss in Section 10, the vanishing of almost principal minors of a Gaussian covariance matrix has the probabilistic meaning of conditional independence [60].

## §10. Conditional Independence

Let $X = (X_i : i \in V)$ be a Gaussian random vector with covariance matrix $\Sigma \in PD_V$. Let $i, j \in V$ be two distinct indices, and let $S \subseteq V \setminus \{i, j\}$. The random variables $X_i$ and $X_j$ are conditionally independent given $X_S$ if and only if $\det(\Sigma_{iS\times jS}) = 0$; see [21, Chapter 3] and [48].

It is fully understood which conditional independence relations hold in the covariance matrices of a linear structural equation model. The following concepts are needed to state the result. Let $\pi$ be a semi-walk from node $i$ to node $j$ in the considered mixed graph $G = (V, D, B)$, and let node $k$ be a non-endpoint of $\pi$. A *collider* on $\pi$ is a node $k$ that is an internal node on $\pi$, and a head on the two edges that precede and succeed $k$ on $\pi$. We recall our convention that the two nodes incident to a bidirected edge are both heads. Pictorially, $\pi$ includes the segment $\to k \leftarrow$, $\to k \leftrightarrow$, $\leftrightarrow k \leftarrow$ or $\leftrightarrow k \leftrightarrow$. A *non-collider* on $\pi$ is an internal node of $\pi$ that is not a collider on $\pi$.

**Definition 10.1.** *Fix a set $S \subseteq V$. Two nodes $i, j \in V$ are* d-connected *by $S$ if $G$ contains a semi-walk from $i$ to $j$ that has all colliders in $S$ and all non-colliders outside $S$. If no such semi-walk exists, then $i$ and $j$ are* d-separated *by $S$.*

The following theorem was first proven for acyclic digraphs [39] and later extended to cover arbitrary mixed graphs [58].

**Theorem 10.2.** *Let $i$ and $j$ be two distinct nodes of a mixed graph $G = (V, D, B)$, and let $S \subseteq V \setminus \{i, j\}$. Then* $\det(\phi_G(\Lambda, \Omega)_{iS\times jS}) = 0$ *for all $\Lambda \in \mathbb{R}^D_{\text{reg}}$, $\Omega \in PD(B)$ if and only if $i$ and $j$ are d-separated by $S$.*

For acyclic digraphs, the theorem can be derived in a probabilistic setup that extends to non-Gaussian conditional independence [46]. Examples with three binary variables, such as the one in Figure 10.1, serve well to illustrate the intuition behind d-separation. In our linear Gaussian setting, Theorem 10.2 is a special case of Theorem 11.3 that we treat in more detail.

Define the *conditional independence ideal*

$$\mathcal{I}_{\text{CI}}(G) \;=\; \langle\, \det(\Sigma_{iS\times jS}) \;:\; i, j \text{ d-separated by } S \,\rangle.$$

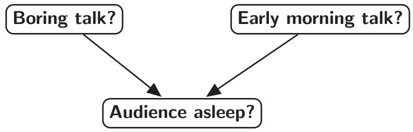


Fig. 10.1. A graph for three binary variables.

By Theorem 10.2, $\mathcal{I}_{\mathrm{CI}}(G) \subseteq \mathcal{I}(G)$ for any mixed graph $G$. In Example 9.1, $\mathcal{I}_{\mathrm{CI}}(G) = \mathcal{I}(G)$ but this may be false even for acyclic digraphs [61]. Nevertheless, if $G$ is an acyclic digraph, then the set of covariance matrices $\mathcal{M}_G$ is cut out by conditional independence relations, that is,

$$\mathcal{M}_G \;=\; V\left(\mathcal{I}_{\mathrm{CI}}(G)\right) \cap PD_V. \tag{10.1}$$

Here, $V(\mathcal{I}_{\mathrm{CI}}(G))$ is the variety of the conditional independence ideal, i.e., the set of symmetric matrices that have all the determinants generating $\mathcal{I}_{\mathrm{CI}}(G)$ zero. The equality in (10.1) also holds when $G$ is an ancestral graph, as defined in Section 7. However, it is false in general as can be seen in Example 3.14. We remark that for acyclic digraphs it has been proven that saturating the conditional independence with respect to all principal minors yields the vanishing ideal [52]:

$$\mathcal{I}(G) \;=\; \mathcal{I}^{\mathrm{CI}}(G) \;:\; \left(\prod_{A \subset V} \det(\Sigma_{A\times A})\right)^{\infty}.$$

For acyclic digraphs and more generally ancestral graphs, the equality from (10.1) allows us to answer Question 3.7 on covariance equivalence by checking whether two graphs have the same d-separation relations. The latter comparison can be accomplished in polynomial time. We state the result for acyclic digraphs [35, 69]. A generalization for ancestral graphs was given more recently [1]; see also [75].

**Theorem 10.3.** *Let $G$ and $G'$ be two acyclic digraphs. Then $\mathcal{M}_G = \mathcal{M}_{G'}$ if and only if $G$ and $G'$ have the same adjacencies and the same unshielded colliders. An unshielded collider is an induced subgraph of the form $i \to j \leftarrow k$.*

We remark that it can also be decided in polynomial time whether two digraphs that are not necessarily acyclic have the same d-separation relations [51]. When the graphs have directed cycles then d-separation equivalence is necessary but not sufficient for covariance equivalence; see, e.g., the example in [15].

## §11. Trek-Separation

We now turn to the characterization of the vanishing of general minors of the covariance matrices in a linear structural equation model. Our first example clarifies the importance of minors that are not almost principal.

**Example 11.1.** If $G$ is the graph from Figure 3.2 and Example 3.14, then $\mathcal{I}(G)$ is generated by $\det(\Sigma_{12,34})$. Such off-diagonal $2 \times 2$ minors are known as tetrads in the statistical literature; see e.g. [20].

The tetrad representation theorem gives a combinatorial characterization of the vanishing of any $2 \times 2$ determinant [57]. The theorem has been greatly generalized, and we now have a full combinatorial understanding of when a minor of the covariance matrix vanishes based on the notion of trek-separation [64]. Moreover, the non-vanishing determinants can be described precisely [14].

**Definition 11.2.** *Let $A, C, S_A, S_C \subseteq V$ be subsets of the vertex set of the mixed graph $G = (V, D, B)$. The pair $(S_A, S_C)$* trek-separates *$A$ and $C$ if every trek from $i \in A$ to $j \in C$ intersects $S_A$ on its left side or $S_C$ on its right side.*

**Theorem 11.3.** *Let $G$ be a mixed graph. Then the $A \times C$ submatrix of $\phi_G(\Lambda, \Omega)$ has generic rank $\leq r$ if and only if there exist sets $S_A$ and $S_C$ with $|S_A| + |S_C| \leq r$ such that $(S_A, S_C)$ trek-separates $A$ and $C$.*

The theorem for acyclic mixed graphs is proven in [64]. The cases with directed cycles are covered by the results in [14]. We describe the ideas behind the proof.

*Proof outline.* The problem can be reduced to the case where $G$ is a digraph by subdividing bidirected edges. For each edge $i \leftrightarrow j \in B$ we introduce a new node $\{i, j\}$ and two edges $\{i, j\} \to i$ and $\{i, j\} \to j$. The new digraph $G'$ thus has $|V| + |B|$ nodes and $|D| + 2|B|$ edges. For instance, if $G$ is the graph from Figure 1.2, then $G'$ is the digraph in Figure 1.1. Every trek in $G$ corresponds to a trek in $G'$ where an edge $i \leftrightarrow j$ in $G$ corresponds to $i \leftarrow \{i, j\} \to j$ in $G'$. The entries of $(\phi_G)_{A,C}$ and those of $(\phi_{G'})_{A,C}$ can be shown to admit the same set of relations.

In the sequel, assume that $G$ itself is a digraph. Then $PD(B)$ contains diagonal matrices with positive entries. The rank of a submatrix of $\phi_G(\Lambda, \Omega)$ for $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega \in PD(B)$ is then the same as the rank of the corresponding submatrix of $\Sigma = \phi_G(\Lambda, I)$.

To establish the claim, we may study the vanishing of the determinants of square submatrices. So assume that $|A| = |C| = r + 1$. The

Cauchy-Binet formula gives that

(11.1)
$$\det\left(\Sigma_{A\times C}\right) = \sum_E \det\left(\left[(I-\Lambda)^{-1}\right]_{E\times A}\right)\det\left(\left[(I-\Lambda)^{-1}\right]_{E\times C}\right),$$

where the sum is over subsets $E \subseteq V$ of cardinality $r+1$. By the Lindström-Gessel-Viennot lemma for general digraphs, it holds that

$$\det\left(\left[(I-\Lambda)^{-1}\right]_{E\times A}\right) = 0 \quad \text{for all } \Lambda \in \mathbb{R}^D_{\text{reg}}, \tag{11.2}$$

if and only if no system of $r+1$ directed paths from $E$ to $A$ is vertex-disjoint. Applying this to all terms in (11.1) shows that $\det(\Sigma_{A\times C})$ vanishes if and only if every system of treks from $A$ to $C$ has a sided intersection. Here, an intersection on the left side of a trek corresponds to the vanishing of determinants of the matrices $\left[(I-\Lambda)^{-1}\right]_{E\times A}$ and, similarly, intersections on the right side are related to the determinants of the matrices $\left[(I-\Lambda)^{-1}\right]_{E\times C}$. The characterization by sided intersections in trek systems can be turned into the claimed statement about trek-separation via Menger's theorem. To account for the distinct role played by the left and the right sides of the treks, Menger's theorem is applied in a digraph $\tilde{G}$ that results from duplicating the nodes and edges of $G$. Each node and each edge of $G$ has a left and a right side version in $\tilde{G}$. Menger's theorem yields a cut set $S$ of cardinality $|S| \le r$ in $\tilde{G}$. Partitioning $S$ according to the left and right side yields the pair $(S_A, S_C)$ for the claimed trek-separation. Q.E.D.

**Example 11.4.** When $G$ is the graph from Figure 3.2, then the submatrix $[\phi_G(\Lambda,\Omega)]_{12,34}$ has generic rank 1; recall Example 11.1. Theorem 11.3 shows that the rank is at least 1 because $(\emptyset,\emptyset)$ does not trek-separate $\{1,2\}$ and $\{3,4\}$. For instance, there is the trek $1\to 3$. That the rank is no larger than 1 follows from $(\emptyset,\{3\})$ trek-separating $\{1,2\}$ and $\{3,4\}$. For instance, the node 3 is on the right side of the two treks $1\to 3$ and $2\to 3\to 4$.

**Example 11.5.** What is the generic rank of the $A\times C$ submatrix of $\phi_G(\Lambda,\Omega)$ when $G$ is the 'spider graph' from Figure 11.1, $A=\{1,2,3,4\}$ and $C=\{5,6,7\}$? Clearly, node $c$ is on every trek between $A$ and $C$. Hence, $(\{c\},\{c\})$ trek-separates $A$ and $C$ and the rank is at most two. In fact, the rank is two. Consider the two treks

$$\pi_1 : 1 \leftrightarrow \circ \to c \to 5,$$
$$\pi_2 : 3 \leftarrow c \leftarrow \circ \leftrightarrow 6.$$

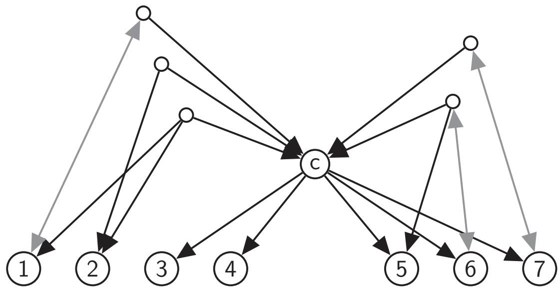


Fig. 11.1. A 'spider graph' with $\{1,2,3,4\}\times\{5,6,7\}$ submatrix of rank two.

They have only node $c$ in common but $c \in \text{right}(\pi_1)$ and $c \in \text{left}(\pi_2)$. Hence, a pair of trek-separating sets must use at least two nodes.

Trek-separation solves the problem of characterizing the vanishing of determinants. However, there is currently no efficient method for deciding when two mixed graphs are trek-separation equivalent.

## §12. Verma Constraints

Much interesting ground lies beyond determinants of the covariance matrix. We content ourselves with two examples concerning a relation first presented in [69].

**Example 12.1.** Let $G$ be the graph from Figure 2.1. The graph has no trek-separation relations as can be checked with the package 'GraphicalModels' for Macaulay2 [37]. The commands

```
needsPackage "GraphicalModels";
G = mixedGraph(digraph {{1,{2,3}},{2,{3}},{3,{4}}}, bigraph {{2,4}});
R = gaussianRing G;
trekIdeal(R,G)
```

return the zero ideal. Issuing the command

```
gaussianVanishingIdeal R
```

shows that $\mathcal{I}(G)$ is generated by

(12.1)

$$\begin{aligned} f_{\text{Verma}} = {} & \sigma_{11}\sigma_{13}\sigma_{22}\sigma_{34} - \sigma_{11}\sigma_{13}\sigma_{23}\sigma_{24} - \sigma_{11}\sigma_{14}\sigma_{22}\sigma_{33} + \sigma_{11}\sigma_{14}\sigma_{23}^2 \\ & - \sigma_{12}^2\sigma_{13}\sigma_{34} + \sigma_{12}^2\sigma_{14}\sigma_{33} + \sigma_{12}\sigma_{13}^2\sigma_{24} - \sigma_{12}\sigma_{13}\sigma_{14}\sigma_{23}. \end{aligned}$$

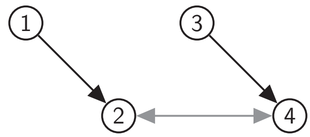


Fig. 12.1. The largest mixed component of the Verma graph.

Clearly, $f_{\mathrm{Verma}}$ is not a determinant. Its vanishing can be explained as follows. Decompose $G$ into its mixed components. The largest component $G[\{2,4\}]$ is depicted in Figure 12.1. By Theorem 6.4, there is a rational function $\tau_{\{2,4\}} : \mathcal{M}_G \to \mathcal{M}_{G'}$, so $\tau_{\{2,4\}}$ is the covariance matrix for $G[\{2,4\}]$. In $G[\{2,4\}]$, there is no trek from node 1 to node 4. Hence, the $(1,4)$ entry of $\tau_{\{2,4\}}(\Sigma)$ is zero. Clearing the denominator yields $f_{\mathrm{Verma}}(\Sigma) = 0$ for $\Sigma \in \mathcal{M}_G$.

The example suggests that new relations can be discovered by decomposing the graph and studying d-separation or trek-separation relations in the mixed components. However, the matter is not as simple as applying a single decomposition.

**Example 12.2.** In order to prevent decomposition of the Verma graph, add a fifth node and bidirected edges such that $(V, B)$ becomes connected. Specifically, consider the graph $G$ from Figure 12.2. The new graph $G$ is simple and acyclic and, thus, $\mathcal{M}_G$ has expected dimension 13. Because the nodes 2 and 5 are d-separated by node 1, we have $\sigma_{12}\sigma_{15} - \sigma_{11}\sigma_{25} \in \mathcal{I}(G)$. Other relations must exist, and indeed a Gröbner basis computation shows that

$$\mathcal{I}(G) \;=\; \langle \sigma_{12}\sigma_{15} - \sigma_{11}\sigma_{25},\; f_{\mathrm{Verma}} \rangle : \sigma_{11}^{\infty}$$

with $f_{\mathrm{Verma}}$ being the polynomial from (12.1). The fact that $f_{\mathrm{Verma}} \in \mathcal{I}(G)$ is explained by Theorem 5.2. Since no directed edge of $G$ has a tail at node 5, the theorem allows us to consider the subgraph induced by the set of nodes $\{1,2,3,4\}$. We are back in Example 12.1 and decomposition yields $f_{\mathrm{Verma}}$ as a relation.

It is clear that more contrived examples can be constructed in which d-/trek-separation applies only after several rounds of alternating between graph decomposition and forming a subgraph induced by an ancestral set. An overview of what is known about such a recursive approach can be found in [55], where the focus is on non-linear models and manipulation of probability density functions.

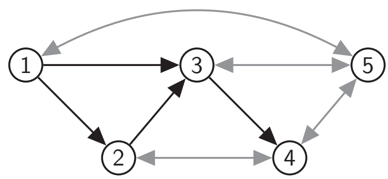


Fig. 12.2. A graph that can be decomposed only after removing the sink node 5.

## §13. Conclusion

Linear structural equation models have covariance matrices with rich algebraic structure. As we showed in Section 3, statistical considerations motivate a host of different algebraic problems. In this review, we focused on methods for parameter identification as well as relations among covariances. While much progress has been made, and continues to be made [30], we have encountered a plethora of open problems of algebraic and combinatorial nature.

In our review, we focused exclusively on linear and Gaussian models. As noted repeatedly, many of the questions have interesting generalizations to non-linear or non-Gaussian models. In particular, in settings with discrete random variables, as considered in [27, 28], algebra and Gröbner basis techniques continue to be useful [21]. Similarly, many additional challenges arise in models with explicit latent variables, which motivate studying maps that result from projecting onto a principal submatrix of $\phi_G$ (recall Section 5).

## References

[ 1 ] R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes, *Markov equivalence for ancestral graphs*, Ann. Statist. **37** (2009), no. 5B, 2808–2837.

[ 2 ] Theodore W. Anderson, *An introduction to multivariate statistical analysis*, third ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. MR 1990662

[ 3 ] Vladimir I. Arnol′d, Sabir M. Guseĭn-Zade, and Aleksandr N. Varchenko, *Singularities of differentiable maps. Vol. II*, Monographs in Mathematics, vol. 83, Birkhäuser Boston, Inc., Boston, MA, 1988, Monodromy and asymptotics of integrals, Translated from the Russian by Hugh Porteous, Translation revised by the authors and James Montaldi. MR 966191

[ 4 ] Rina Foygel Barber, Mathias Drton, and Luca Weihs, *Semid: Identifiability of linear structural equation models*, 2015, R package version 0.2.

[ 5 ] Kenneth A. Bollen, *Structural equations with latent variables*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1989, A Wiley-Interscience Publication. MR 996025 (90k:62001)

[ 6 ] Kenneth A. Bollen and Kwok-Fai Ting, *A tetrad test for causal indicators*, Psychological methods **5** (2000), no. 1, 3–22.

[ 7 ] Carlos Brito and Judea Pearl, *A new identification condition for recursive models with correlated errors*, Struct. Equ. Model. **9** (2002), no. 4, 459–474. MR 1930449

[ 8 ] Andries E. Brouwer and Jan Draisma, *Equivariant Gröbner bases and the Gaussian two-factor model*, Math. Comp. **80** (2011), no. 274, 1123–1133. MR 2772115

[ 9 ] Søren L. Buhl, *On the existence of maximum likelihood estimators for graphical Gaussian models*, Scand. J. Statist. **20** (1993), no. 3, 263–270.

[10] Bryant Chen, *Decomposition and identification of linear structural equation models*, ArXiv e-prints (2015), 1508.01834.

[11] Bryant Chen, Jin Tian, and Judea Pearl, *Testable implications of linear structural equations models*, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (Palo Alto, CA) (Carla E. Brodley and Peter Stone, eds.), AAAI Press, 2014, pp. 2424–2430.

[12] David Cox, John Little, and Donal O'Shea, *Ideals, varieties, and algorithms*, third ed., Undergraduate Texts in Mathematics, Springer, New York, 2007, An introduction to computational algebraic geometry and commutative algebra. MR 2290010 (2007h:13036)

[13] Wolfram Decker, Gert-Martin Greuel, Gerhard Pfister, and Hans Schönemann, SINGULAR *4-0-2 — A computer algebra system for polynomial computations*, `http://www.singular.uni-kl.de`, 2015.

[14] Jan Draisma, Seth Sullivant, and Kelli Talaska, *Positivity for Gaussian graphical models*, Adv. in Appl. Math. **50** (2013), no. 5, 661–674. MR 3044565

[15] Mathias Drton, *Likelihood ratio tests and singularities*, Ann. Statist. **37** (2009), no. 2, 979–1012. MR 2502658

[16] Mathias Drton, Rina Foygel, and Seth Sullivant, *Global identifiability of linear structural equation models*, Ann. Statist. **39** (2011), no. 2, 865–886.

[17] Mathias Drton, Hélène Massam, and Ingram Olkin, *Moments of minors of Wishart matrices*, Ann. Statist. **36** (2008), no. 5, 2261–2283. MR 2458187

[18] Mathias Drton and Martyn Plummer, *A Bayesian information criterion for singular models*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **79** (2017), 323–380.

[19] Mathias Drton and Thomas S Richardson, *Multimodality of the likelihood in the bivariate seemingly unrelated regressions model*, Biometrika **91** (2004), no. 2, 383–392.

[20] Mathias Drton, Bernd Sturmfels, and Seth Sullivant, *Algebraic factor analysis: tetrads, pentads and beyond*, Probab. Theory Related Fields **138** (2007), no. 3-4, 463–493. MR 2299716

[21] ———, *Lectures on algebraic statistics*, Oberwolfach Seminars, vol. 39, Birkhäuser Verlag, Basel, 2009. MR 2723140 (2012d:62004)

[22] Mathias Drton and Luca Weihs, *Generic identifiability of linear structural equation models by ancestor decomposition*, Scand. J. Stat. **43** (2016), no. 4, 1035–1045. MR 3573674

[23] Mathias Drton and Han Xiao, *Finiteness of small factor analysis models*, Ann. Inst. Statist. Math. **62** (2010), no. 4, 775–783. MR 2652316

[24] ———, *Wald tests of singular hypotheses*, Bernoulli **22** (2016), no. 1, 38–59. MR 3449776

[25] Mathias Drton and Josephine Yu, *On a parametrization of positive semidefinite matrices with zeros*, SIAM J. Matrix Anal. Appl. **31** (2010), no. 5, 2665–2680. MR 2740626 (2011j:15057)

[26] Jan Ernest, Dominik Rothenhäusler, and Peter Bühlmann, *Causal inference in partially linear structural equation models: identifiability and estimation*, ArXiv e-prints (2016), 1607.05980.

[27] Robin J. Evans, *Graphs for margins of Bayesian networks*, Scand. J. Stat. **43** (2016), no. 3, 625–648. MR 3543314

[28] Robin J. Evans and Thomas S. Richardson, *Smooth, identifiable supermodels of discrete DAG models with latent variables*, ArXiv e-prints (2015), 1511.06813.

[29] William N. Evans and Jeanne S. Ringel, *Can higher cigarette taxes improve birth outcomes?*, Journal of Public Economics **72** (1999), no. 1, 135–154.

[30] Alex Fink, Jenna Rajchgot, and Seth Sullivant, *Matrix Schubert varieties and Gaussian conditional independence models*, J. Algebraic Combin. **44** (2016), no. 4, 1009–1046. MR 3566228

[31] Franklin M. Fisher, *A correspondence principle for simultaneous equation models*, Econometrica **38** (1970), 73–92.

[32] Christopher Fox, *Interpretation and inference of linear structural equation models*, Ph.D. thesis, University of Chicago, 2014.

[33] Christopher J. Fox, Andreas Käufl, and Mathias Drton, *On the causal interpretation of acyclic mixed graphs under multivariate normality*, Linear Algebra Appl. **473** (2015), 93–113. MR 3338327

[34] Rina Foygel, Jan Draisma, and Mathias Drton, *Half-trek criterion for generic identifiability of linear structural equation models*, Ann. Statist. **40** (2012), no. 3, 1682–1713.

[35] Morten Frydenberg, *The chain graph Markov property*, Scand. J. Statist. **17** (1990), no. 4, 333–353. MR 1096723

[36] Luis D. Garcia-Puente, Sarah Spielvogel, and Seth Sullivant, *Identifying causal effects with computer algebra*, Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI) (Peter Grünwald and Peter Spirtes, eds.), AUAI Press, 2010.

[37] Luis David García-Puente, Sonja Petrović, and Seth Sullivant, *Graphical models*, J. Softw. Algebra Geom. **5** (2013), 1–7. MR 3073716
[38] Dan Geiger, David Heckerman, Henry King, and Christopher Meek, *Stratified exponential families: graphical models and model selection*, Ann. Statist. **29** (2001), no. 2, 505–529. MR 1863967
[39] Dan Geiger, Thomas Verma, and Judea Pearl, *Identifying independence in Bayesian networks*, Networks **20** (1990), no. 5, 507–534, Special issue on influence diagrams. MR 1064736
[40] Daniel R. Grayson and Michael E. Stillman, *Macaulay2, a software system for research in algebraic geometry*, Available at http://www.math.uiuc.edu/Macaulay2/.
[41] Frank Harary, *The determinant of the adjacency matrix of a graph*, SIAM Rev. **4** (1962), 202–210. MR 0144330
[42] Dominique M. A. Haughton, *On the choice of a model to fit data from an exponential family*, Ann. Statist. **16** (1988), no. 1, 342–355. MR 924875
[43] Alain Hauser and Peter Bühlmann, *Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **77** (2015), no. 1, 291–318. MR 3299409
[44] Jan T. A. Koster, *Marginalizing and conditioning in graphical models*, Bernoulli **8** (2002), no. 6, 817–840. MR 1963663
[45] Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik Hoyer, *Discovering cyclic causal models by independent components analysis*, Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08) (Corvallis, Oregon), AUAI Press, 2008, pp. 366–374.
[46] Steffen L. Lauritzen, *Graphical models*, Oxford Statistical Science Series, vol. 17, The Clarendon Press, Oxford University Press, New York, 1996, Oxford Science Publications. MR 1419991
[47] Dennis Leung, Mathias Drton, and Hisayuki Hara, *Identifiability of directed Gaussian graphical models with one latent source*, Electron. J. Stat. **10** (2016), no. 1, 394–422. MR 3466188
[48] Radim Lněnička and František Matúš, *On Gaussian conditional independent structures*, Kybernetika (Prague) **43** (2007), no. 3, 327–342. MR 2362722
[49] Judea Pearl, *Causality*, second ed., Cambridge University Press, Cambridge, 2009, Models, reasoning, and inference. MR 2548166 (2010i:68148)
[50] Thomas Richardson and Peter Spirtes, *Ancestral graph Markov models*, Ann. Statist. **30** (2002), no. 4, 962–1030. MR 1926166 (2003h:60017)
[51] Thomas S. Richardson, *A characterization of Markov equivalence for directed cyclic graphs*, International Journal of Approximate Reasoning **17** (1997), no. 2/3, 107–162.
[52] Hajir Roozbehani and Yury Polyanskiy, *Algebraic methods of classifying directed graphical models*, 2014 IEEE International Symposium on Information Theory, June 2014, pp. 2027–2031.

[53] N. Shiers, P. Zwiernik, J. A. D. Aston, and J. Q. Smith, *The correlation space of Gaussian latent tree models and model selection without fitting*, Biometrika **103** (2016), no. 3, 531–545. MR 3551782

[54] Shohei Shimizu, *LiNGAM: Non-Gaussian methods for estimating causal structures*, Behaviormetrika **41** (2014), no. 1, 65–98.

[55] Ilya Shpitser, Robin Evans, Thomas Richardson, and James Robins, *Introduction to nested Markov models*, Behaviormetrika **41** (2014), no. 1, 3–39.

[56] Ilya Shpitser and Judea Pearl, *Identification of joint interventional distributions in recursive semi-Markovian causal models*, Proceedings of the Twenty-First National Conference on Artificial Intelligence (Menlo Park, CA), AAAI Press, 2006, pp. 1219–1226.

[57] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, prediction, and search*, second ed., Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2000, With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book. MR 1815675 (2001j:62009)

[58] Peter Spirtes, Thomas Richardson, Chris Meek, Richard Scheines, and Clark Glymour, *Using path diagrams as a structural equation modelling tool*, Sociological Methods and Research **27** (1998), 182–225.

[59] Paul D. Stolley, *When genius errs: R. A. Fisher and the lung cancer controversy*, Am. J. Epidemiol. **133** (1991), no. 5, 416–425.

[60] Milan Studený, *Probabilistic conditional independence structures*, Information Science and Statistics, Springer, London, 2005. MR 3183760

[61] Seth Sullivant, *Algebraic geometry of Gaussian Bayesian networks*, Adv. in Appl. Math. **40** (2008), no. 4, 482–513. MR 2412156

[62] ———, *A Gröbner basis for the secant ideal of the second hypersimplex*, J. Commut. Algebra **1** (2009), no. 2, 327–338. MR 2504939

[63] Seth Sullivant and Elizabeth Gross, *The maximum likelihood threshold of a graph*, Bernoulli **24** (2018), no. 1, 386–407.

[64] Seth Sullivant, Kelli Talaska, and Jan Draisma, *Trek separation for Gaussian graphical models*, Ann. Statist. **38** (2010), no. 3, 1665–1685. MR 2662356 (2011f:62076)

[65] Jin Tian, *Identifying direct causal effects in linear models*, Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI), 2005, pp. 346–353.

[66] Jin Tian and Judea Pearl, *A general identification condition for causal effects*, AAAI/IAAI, 2002, pp. 567–573.

[67] Caroline Uhler, *Geometry of maximum likelihood estimation in Gaussian graphical models*, Ann. Statist. **40** (2012), no. 1, 238–261. MR 3014306

[68] Lieven Vandenberghe and Martin S. Andersen, *Chordal graphs and semidefinite optimization*, Foundations and Trends in Optimization **1** (2014), no. 4, 241–433.

[69] Thomas S. Verma and Judea Pearl, *Equivalence and synthesis of causal models*, Uncertainty in Artificial Intelligence 6, Elsevier, 1991, UCLA Cognitive Systems Laboratory, Technical Report (R-150), pp. 255–268.

[70] Sumio Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge Monographs on Applied and Computational Mathematics, vol. 25, Cambridge University Press, Cambridge, 2009. MR 2554932

[71] Luca Weihs, Bill Robinson, Emilie Dufresne, Jennifer Kenkel, Kaie Kubjas, Reginald McGee, II, Nhan Nguyen, Elina Robeva, and Mathias Drton, *Determinantal generalizations of instrumental variables*, Journal of Causal Inference **6** (2018), no 1.

[72] Nanny Wermuth, *Probability distributions with summary graph structure*, Bernoulli **17** (2011), no. 3, 845–879. MR 2817608

[73] Sewall Wright, *Correlation and causation*, J. Agricultural Research **20** (1921), 557–585.

[74] ———, *The method of path coefficients*, Ann. Math. Statist. **5** (1934), no. 3, 161–215.

[75] Hui Zhao, Zhongguo Zheng, and Baijun Liu, *On the Markov equivalence of maximal ancestral graphs*, Sci. China Ser. A **48** (2005), no. 4, 548–562. MR 2157690 (2006e:62069)

*Department of Statistics, University of Washington, Seattle, WA 98195-4322, U.S.A.*
*E-mail address*: `md5@uw.edu`