Tractable Structure Search in the Presence of Latent Variables

Thomas Richardson, Heiko Bailer, Moulinath Banerjee
Department of Statistics, University of Washington
{tsr, heiko, mouli} @ stat.washington.edu

Abstract

The problem of learning the structure of a DAG model in the presence of latent variables presents many formidable challenges. In particular there are an infinite number of latent variable models to consider, and these models possess features which make them hard to work with. We describe a class of graphical models which can represent the conditional independence structure induced by a latent variable model over the observed margin. We give a parametrization of the set of Gaussian distributions with conditional independence structure given by a MAG model. The models are illustrated via a simple example. Different estimation techniques are discussed in the context of Zellner's Seemingly Unrelated Regression (SUR) models.

Keywords: Multivariate Graphical Models; Causal Modelling; Latent Variables; Ancestral Graphs; MAG Models.

1 INTRODUCTION

There has been significant progress in the development of algorithms for learning the directed acyclic graph (DAG) part of a Bayesian network from complete data and optional background knowledge (Cooper and Herskovits, 1992; Spirtes, Glymour and Scheines, 1993; Heckerman, Geiger and Chickering, 1994). However, the problem of learning the DAG part of a Bayesian network with latent (unmeasured) variables is more difficult: first the number of possible models is infinite, and second, calculating scores for latent variable (LV) models is generally much slower than calculating scores for models without LVs. In addition, general LV models have a number of other features which make them hard to work with: LV models may be overparametrized, e.g. containing edges or nodes that are redundant, and as a consequence the parameters may be underidentified, leading to multimodal or flat likelihood surfaces; further, as described by Geiger and Meek (1998), LV models are stratified exponential families, rather than curved exponential families (like DAGs without LVs), and consequently the results which guarantee the asymptotic consistency of scores such as BIC do not apply; finally, LV models do not have a well-defined dimension.

This presents a dilemma: on the one hand, attempting to search for causal structure without allowing for the possibility of latent or missing variables is substantively unreasonable in many contexts, and yet on the other hand, explicitly including latent variables appears to make the search space intractable, and introduces models with features that make model selection difficult. To address this problem Spirtes, Meek and Richardson (1997) have introduced a class of graphical Gaussian models, called MAG models, which do not include latent variables, but do impose the independence constraints given by latent variable models, and *only* these constraints. Thus for any given LV model there is a MAG model to which it is Markov equivalent. However, since LV models often impose non-independence constraints, the corresponding MAG model will parametrize a set of distributions which form a superset of the distribution parametrized by the LV model. In contrast to latent variable models, MAG models are efficiently parametrized, always statistically identifiable, and have a well-defined dimension (they form curved exponential families).

We describe a class of graphical models which can represent the conditional independence structure induced by a latent variable model over the observed margin. We give a parametrization of the set of Gaussian distributions with conditional independence structure given by a MAG model. The models are illustrated via a simple example taken from Whittaker (1990). We discuss the relationship between Seemingly Unre-

lated Regression (SUR) models (Zellner, 1962). Finally, estimation techniques for MAGs, based on SUR methods are described. The performance of two estimation methods are compared on the example taken from Whittaker.

2 ANCESTRAL GRAPHS

A *mixedgraph* is a graph containing three types of edge $\{-, \rightarrow, \leftrightarrow\}$, where at most one edge (of any type) may occur between each pair of vertices.

We naturally extend Pearl's d-separation criterion to mixed graphs as follows: a pair of consecutive edges meeting at a vertex z on a path form a *collider* if both edges have an arrowhead at z, i.e. $\rightarrow z \leftarrow$, $\leftrightarrow z \leftrightarrow$, $\leftrightarrow z \leftarrow$, $\rightarrow z \leftrightarrow$. Two consecutive edges which do not form a collider are said to form a *non-collider*. A vertex a is said to be an *ancestor* of a vertex b if **either** there is a directed path $a \rightarrow \cdots \rightarrow b$ on which every edge is of the form '$\rightarrow$', and has the same orientation, **or** $a = b$.

A path between vertices x and y in a mixed graph is said to be *d-connecting given a set* $\mathbf{Z}$ if

(i) every non-collider on the path is not in $\mathbf{Z}$, and

(ii) every collider on the path is an ancestor of $\mathbf{Z}$, (note: each vertex is its own ancestor).

If there is no path d-connecting x and y given $\mathbf{Z}$, then x and y are said to be d-separated given $\mathbf{Z}$. Sets $\mathbf{X}$ and $\mathbf{Y}$ are said to be *d-separated* given $\mathbf{Z}$,if for every pair x, y, with $x \in \mathbf{X}$ and $y \in \mathbf{Y}$, x and y are d-separated given $\mathbf{Z}$.

An *ancestral* mixed graph, is a mixed graph satisfying the following conditions:

(i) There are no directed cycles: if there is a directed path from a to b, i.e. $a \rightarrow \ldots \rightarrow b$, then there is no directed path from b to a.

(ii) If there is an edge $a \leftrightarrow b$ in the graph, then there is no directed path from b to a, and there is no directed path from a to b.

(iii) If there is an edge $a - b$ in the graph, then there is no edge with an arrowhead at a, i.e. there is no edge $c \rightarrow a$, $c \leftrightarrow a$.

A graph satisfying these properties is termed 'ancestral' because whenever there is an arrowhead at a, i.e. $a \leftarrow b$ or $a \leftrightarrow b$ then a is **not** an ancestor of b in the graph. The presence of a tail at a, i.e. $a \rightarrow b$ or $a - b$ can also be given an ancestral interpretation in the context of a larger graph containing selection variables (see Section 3).

Figure 1 (a) shows two mixed graphs that are not ancestral: in the first there is a directed path from b to a while at the same time there is an edge $a \leftrightarrow b$; in the second there is an edge $a - c$ while at the same time there is an edge $a \leftarrow b$. Figure 1 (b) shows two ancestral mixed graphs. Note that the class of ancestral mixed graphs contains the set of DAGs and undirected graphs. However, as the second graph in Figure 1(a) shows, the class of ancestral mixed graphs does not contain the class of chain graphs.

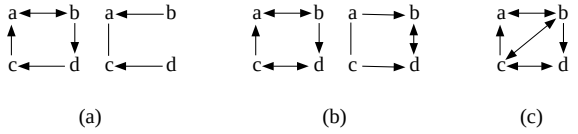

Figure 1: (a) Two mixed graphs that are not ancestral; (b) two ancestral mixed graphs; (c) an ancestral mixed graph that is not maximal.

2.1 MAXIMAL MIXED ANCESTRAL GRAPHS (MAGs).

An ancestral mixed graph is said to be *maximal* if it satisfies the following condition:

(iv) If there is no edge between a and b, then there is some subset $\mathbf{Z}$ of the other vertices such that a and b are d-separated given $\mathbf{Z}$.

Figure 1(c) shows a simple example of an ancestral mixed graph which is *not* maximal: there is no edge between a and d in the graph, and yet a and d are d-connected by all four subsets of the other variables $\{b, c\}$. The local Markov properties for DAGs and undirected graphs (see Lauritzen, 1996), imply that every graph in either of these classes is maximal, since there is some subset which d-separates each pair of edges which are not joined by an edge.

Spirtes and Richardson (1997) present an $O(n^4)$ algorithm for checking that a MAG is maximal. The algorithm is based on the following pairwise Markov property which holds in maximal ancestral mixed graphs. A vertex x is said to be *anterior* to y if there is a path from x to y on which every edge $\langle u, v \rangle$ is either undirected, $u - v$, or directed, $u \rightarrow v$, with orientation from x to y. It follows from condition (iii) in the definition of an ancestral graph that if $u \rightarrow v$ is a directed edge then no undirected edge occurs between v and y, so the path is of the form $x - \cdots - \rightarrow \cdots \rightarrow y$, $x - \cdots - y$, or $x \rightarrow \cdots \rightarrow y$. $\mathrm{ant}_{\mathcal{M}}(x)$ is the set of vertices anterior to x in $\mathcal{M}$, and similarly $\mathrm{ant}_{\mathcal{M}}(\mathbf{X}) = \{y \mid y \in \mathrm{ant}_{\mathcal{M}}(x) \text{ for some } x \in \mathbf{X}\}$. Finally note that for a DAG $\mathcal{G}$, x is an anterior to y if and only if x is an ancestor of y.

Pairwise Markov Property:
If there is no edge between a and b in the **maximal** ancestral graph M then:
a is d-separated from b by $(\text{ant}_{\mathcal{M}}(a) \cup \text{ant}_{\mathcal{M}}(b)) \setminus \{a, b\}$

Except where otherwise noted, we will restrict ourselves to Maximal Ancestral mixed Graphs (MAGs). Given given a non-maximal ancestral mixed graph it can be converted into a maximal ancestral graph by adding double-headed edges ($\leftrightarrow$), between every pair of vertices for which no d-separating set exists. Further the resulting maximal ancestral graph will represent exactly the same set of d-separation as held in the original graph (see section 3). In addition, maximal ancestral graphs lead to a natural parametrization of the associated Markov model in the Gaussian case (see section 4).

3 GRAPHS WITH LATENT AND SELECTION VARIABLES

Cox and Wermuth (1996) and Spirtes *et al.* (1997) consider a DAG $\mathcal{G}$ with vertex set $\mathbf{V}$, partitioned into observed ($\mathbf{O}$), latent ($\mathbf{L}$), and selection ($\mathbf{S}$) subsets. The interpretation is that $\mathcal{G}$ represents a causal, or data-generating mechanism; $\mathbf{O}$ represents the subset of the variables that are observed; $\mathbf{S}$ represents a set of variables which, due to the nature of the mechanism selecting the sample, are conditioned on in the subpopulation from which the sample is drawn; the variables $\mathbf{L}$ are not observed and for this reason are called *latent*.

Spirtes *et al.* show that given such a DAG $\mathcal{G}$, with vertex set $\mathbf{V}$, partitioned into $(\mathbf{O}, \mathbf{S}, \mathbf{L})$ there is a corresponding MAG $\mathcal{M}$ with vertex set $\mathbf{O}$, such that for disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}$, $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z} \cup \mathbf{S}$ in $\mathcal{G}$, if and only if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$ in $\mathcal{M}$. Thus the MAG captures the independencies holding among the observed variables in the selected subpopulation.

The algorithm for creating a MAG $\mathcal{M}$ from a DAG $\mathcal{G}$. Requires only three steps:

DAG to MAG Algorithm

(i) Form an undirected graph $\mathcal{M}$ with vertex set $\mathbf{O}$ in which there is an edge $x - y$ if and only if for every subset $\mathbf{Z} \subseteq \mathbf{O}$, x and y are d-connected given $\mathbf{Z} \cup \mathbf{S}$ in $\mathcal{G}$.

(ii) If there is an edge $x - y$ in $\mathcal{M}$, and $x \notin \text{ant}_{\mathcal{G}}(\{y\} \cup \mathbf{S})$, and $y \notin \text{ant}_{\mathcal{G}}(\{x\} \cup \mathbf{S})$ then replace $x - y$ with $x \leftrightarrow y$.

(iii) If there is an edge $x - y$ in $\mathcal{M}$, and $x \in \text{ant}_{\mathcal{G}}(\{y\} \cup \mathbf{S})$, but $y \notin \text{ant}_{\mathcal{G}}(\{x\} \cup \mathbf{S})$ then replace $x - y$ with $x \leftarrow y$.

Step (i), together with the fact stated above, that $\mathcal{M}$ captures all d-separation relations holding between disjoint subsets $\mathbf{X}$, $\mathbf{Y}$ of $\mathbf{O}$ given a third subset $\mathbf{Z}$ union $\mathbf{S}$, ensures that the resulting graph $\mathcal{M}$ is maximal.

It is simple to check that the orientation rules given in steps (ii) and (iii) result in a mixed graph that is ancestral. If there is an edge $x - y$ in $\mathcal{M}$, after steps (ii) and (iii) then both x and y are in $\text{ant}_{\mathcal{G}}(\mathbf{S})$. Since in a DAG $\mathcal{G}$ there are no undirected edges $\text{ant}_{\mathcal{G}}(\mathbf{X}) = \text{an}_{\mathcal{G}}(\mathbf{X})$, the set of ancestors of $\mathbf{X}$ in $\mathcal{G}$, so both x and y are ancestors of vertices in $\mathbf{S}$ (in $\mathcal{G}$).

The algorithm can in fact be applied directly to an ancestral graph (not simply a DAG) whose vertices are partitioned $\mathbf{O}, \mathbf{S}, \mathbf{L}$, to generate a MAG encoding the d-separation relations holding between vertices in $\mathbf{O}$ given $\mathbf{S}$.

3.1 EXAMPLE

We illustrate the operation of this algorithm with a simple example. Figure 2(a) shows an example of a DAG $\mathcal{G}$, while (b) shows the corresponding MAG, under the partition $\mathbf{O} = \{a, b, c, d, e, f\}$, $\mathbf{S} = \{s\}$, and $\mathbf{L} = \{l_1, l_2\}$.

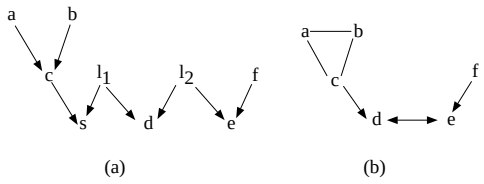

Figure 2: (a) A DAG $\mathcal{G}$ with vertex set $\mathbf{V} = \{a, b, c, d, e, f, l_1, l_2, s\}$; (b) the MAG $\mathcal{M}$ corresponding to $\mathcal{G}$ under the partition $\mathbf{O} = \{a, b, c, d, e, f\}$, $\mathbf{S} = \{s\}$, and $\mathbf{L} = \{l_1, l_2\}$.

3.2 DECOMPOSITION INTO DIRECTED AND UNDIRECTED COMPONENTS

It follows from (i) and (iii) in the definition of an ancestral graph, that it is always possible to construct a total order on the vertices in an ancestral graph in such a way that if a is an ancestor of b then a precedes b in the ordering, *and* such that all vertices x which are endpoints of undirected edges $x - y$ preceed all vertices which are not.

It is thus always possible to partition the vertices of a MAG into two sets Ω, Δ such that the induced subgraph on Ω is completely undirected, and the induced subgraph on Δ contains no undirected edges; further

if there is an edge connecting a vertex $u \in \Omega$ to a vertex $d \in \Delta$ then it is oriented as $u \to d$. (The induced subgraphs on Ω and Δ need not be connected.) A schematic of this decomposition is shown in Figure 3. For the MAG in Figure2, such a partition is $\Omega = \{a, b, c\}$, $\Delta = \{d, e, f\}$. This decomposition is

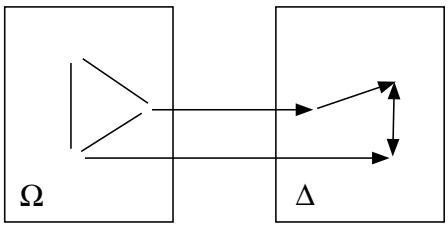

Figure 3: Schematic showing the decomposition of an ancestral mixed graph into an undirected component and a directed component.

significant because it implies a corresponding factorization of the joint density:

$$\mathrm{P}(\Omega, \Delta) = \mathrm{P}(\Omega) \cdot \mathrm{P}(\Delta \mid \Omega)$$

Thus we may break down the problem of estimating a MAG, to the problem of fitting an undirected graphical model, the induced subgraph over Ω, and a MAG containing no undirected edges for the conditional distribution $P(\Delta \mid \Omega)$.

There are well-established methods such as the IPS algorithm (See Lauritzen, 1996) for fitting undirected graphical models. For this reason, in this paper we will focus on the problem of parametrizing, estimating and scoring MAGs containing no undirected edges. This subclass of MAG models is also interesting in its own right since it follows directly from the MAG construction algorithm given in section 3 that if there are no selection variables ($\mathbf{S} =$) then there will be no undirected edges in the resulting MAG.

3.3 RELATION TO SUMMARY GRAPHS

MAGs are closely related to the Summary Graphs in Cox and Wermuth (1996), though there are key differences. In particular it is possible to have no edge between a pair of vertices x and y and yet there is no subset of the remaining variables which make x and y conditionally independent; it is also possible to have more than one edge between a pair of vertices in a summary graph.

Pearl and Verma (1991) prove that every latent variable model may be transformed into a Markov equivalent model in which every latent variable has only two children. However, such latent variable models, though simpler, are not in general characterized purely in terms of conditional independence relations.

4 GAUSSIAN *MAG* MODELS

A Gaussian MAG *model* is a set of multivariate Gaussian distributions P, such that for all disjoint sets $\mathbf{X}$,$\mathbf{Y}$ and $\mathbf{Z}$, if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in the MAG, then $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ in P.

A Gaussian MAG model (with means fixed at zero) can be parametrized as follows:

(i) Associate with each vertex in the MAG a linear equation, expressing that variable as a linear function of its parents plus an error term:

$$y = \sum_{x_i \in \mathrm{pa}(y)} \alpha_i x_i + \epsilon_y$$

where $\mathrm{pa}(y)$ is the set of parents of y.

(ii) Specify a multivariate Gaussian distribution over the error terms (with mean zero) satisfying the condition that if there is not a double-headed edge $X \leftrightarrow Y$ in the MAG, then $\mathrm{Cov}(\epsilon_X, \epsilon_Y) = 0$, but otherwise unrestricted.

The dimension of a MAG model is then equal to the number of vertices plus the number of edges (of either type) that are present in the graph.

Note that the MAG model parametrizes *all* Gaussian distributions in which the conditional independence relations corresponding to d-separation relations hold. For this to hold it is crucial that the ancestral mixed graph be maximal: consider the graph in Figure 1(c). There are no d-separation relations holding in this graph, but under the parametrization given above the corresponding Gaussian model would not be saturated. In particular it implies the following constraint:

$$\mathrm{Cov}(\, a \perp \hat{a}(c), d \perp \hat{d}(b)\,) = 0$$

where $\hat{x}(y)$ is the linear predictor of x from y.

Since the set of distributions parametrized by a Gaussian MAG models is characterized purely in terms of conditional independence it follows that two Gaussian MAG models are Markov equivalent if and only if they are statistically equivalent.

Gaussian MAG models represent a generalization of the Seemingly Unrelated Regression (SUR) models introduced by Zellner (1962). SUR models and associated estimation techniques are discussed in sections 8 & 9. Sewall Wright's path diagrams also contain double headed arrows representing correlated errors (Wright, 1934).

5 EXAMPLE: COCHRAN'S NOCTUID MOTH DATA

To illustrate the use of MAG models we consider the data on moth trappings, which originally appeared in the statistical literature in a paper of Cochran (1938), but which were subsequently analyzed by Dempster (1972) and Whittaker (1990). The data consist of one response variable:

$$moth = \log(1 + \ no.\ of\ moths\ caught\ in\ a\ light\ trap\ on\ one\ night)$$

and five covariates,

min:	the minimum night temperature;
max:	the previous day's maximum temperature;
wind:	the average wind speed during the night;
rain:	the amount of rain during the night;
cloud:	the percentage of starlight obscured by clouds.

Dempster (1972) fitted a model which corresponds to the undirected graph shown in Figure 4(a), in which conditional independence is encoded via separation. Dempster arrived at this model via a forward selection procedure, which terminated with the first model for which the p-value > 0.05, based on a likelihood ratio test against the full model (with d.f. $= 21 \perp$ Dim. of model). We also give $Deviance + \ln(Sample\ \ Size) \cdot Dimension$, since this is equal to the BIC score + a constant (note that lower scores correspond to 'better' models under this criterion).

Whittaker (1990) presents an analysis based on a chain graph with a division of the variables into two blocks, the first containing the five covariates, and the second containing the response, see Figure 4(b).

Applying the FCI search algorithm (described in Spirtes, Glymour and Scheines, 1993) resulted in the MAG shown in Figure 4(c), which imposes the following conditional independence constraints:

$$max \perp\!\!\!\perp rain, cloud, moth;$$

$$min \perp\!\!\!\perp rain, moth \mid cloud;$$

$$wind \perp\!\!\!\perp max, cloud, rain \mid min;$$

$$rain \perp\!\!\!\perp max, min, wind, moth \mid cloud.$$

The FCI algorithm is not designed to maximize the BIC score, but instead uses a sequence of conditional independence tests. However, a subsequent search indicates that there is no MAG model with a higher score. In fact the MAG model is nested within Whittaker's model. Since the two models differ by 2 d.f., and the difference in deviance is only 2.11, a likelihood ratio test finds no evidence against the MAG model (p-value 0.348).

6 MARKOV EQUIVALENCE

The existence of equivalent, yet structurally distinct, DAG models is a central problem when attempting to identify cause-effect relationships (Spirtes, Glymour and Scheines, 1993; Verma and Pearl, 1990). It is important to identify statistically equivalent structures, since they may have different causal interpretations. Data fitting a model M well can be regarded as evidence for some feature of the model only if that feature is common to all models equivalent to M.

Equivalent models may also be important for model search: it is inefficient to score many different structures if it is inevitable that they will all receive the same score. In the case of DAG models without latent variables a number of authors (Madigan *et al.*, 1996; Chickering, 1995; Spirtes and Meek, 1995) have performed model searches which consider equivalence classes of models.

Figure 5 shows three MAGs that are Markov equivalent (hence statistically equivalent) to the MAG in Figure 4(c). In fact it is not hard to show that all MAGs that are statistically equivalent possess the same set of adjacencies. (There are 6 other MAGs in this class).

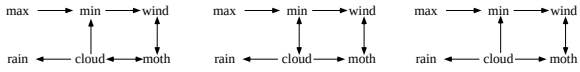

Figure 5: Three MAGs Markov equivalent to the MAG in Figure 4(c).

DAG models with LVs present further problems in this regard since there is no general characterization of the constraints on the marginal distribution over the observed variables that are imposed by LV models. Such a characterization would be needed in order to describe statistical equivalence classes of LV models.

However, the *independence* constraints imposed by latent variable models correspond to the d-separation relations involving only observed variables. It is possible to characterize Markov equivalence classes of LV models in graphical terms (Spirtes, Glymour and Scheines, 1993). A graphical characterization of Markov equivalence (and hence statistical equivalence) for MAGs without undirected edges follows directly from this result. Spirtes and Richardson, (1997) extends these results to cover MAGs including undirected edges.

This graphical characterization can then be exploited to perform searches which consider Markov equivalence classes of MAG models and to identify structural features that are common to all latent variable models

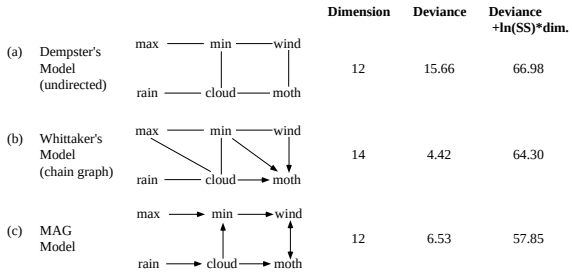

Figure 4: Three graphical models for Cochran's moth data.

within a given Markov equivalence class, parametrized by a MAG. Spirtes, Richardson and Meek (1996) describe such a procedure for model selection that searches equivalence classes of MAGs.

6.1 MARKOV EQUIVALENCE CLASSES ON 3 AND 4 VERTICES

We briefly describe the Markov equivalence classes of MAGs that occur on 3 and 4 variables. This characterization may be important for local search procedures of the type considered by Spirtes and Cooper (1999), which look for models that describe well the relationships between small subsets of variables.

Every MAG on 3 variables is Markov equivalent to some DAG. Thus there are 11 Markov equivalence classes on 3 variables.

Up to permutation of vertices there are 5 classes of MAGs which are not Markov equivalent to any DAG. These are illustrated in Figure 6. Only models in classes (a) and (b) can be decomposed into SUR models, with unconstrained error covariance matrices.

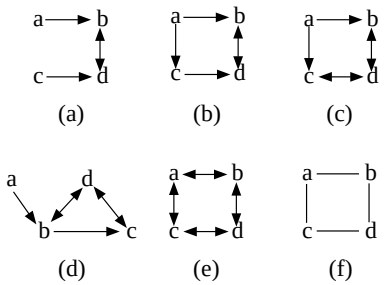

Figure 6: MAGs that are not Markov equivalent to any DAG. The number of Markov equivalence classes of each type are: (a) 12; (b) 12; (c) 12; (d) 3; (e) 24; (f) 3.

Table 1: No. of Markov Equivalence Classes

Class of Graphs	No. of Vertices 3	4
MAGs	11	251
MAGs without undirected edges	11	248
DAGs	11	185
Undirected Graphs	8	64

7 PROCEDURES FOR ESTIMATING GAUSSIAN *MAGS*

Closed form maximum likelihood estimates do not in general exist for MAGs. Thus iterative procedures must be used to find the Maximum Likelihood (ML) estimates. The use of an iterative procedure is undesirable since it will slow down a specification search. However, consistent closed form estimates can be obtained by first regressing each variable on its parents to estimate the linear coefficients, and then using the covariance of the residuals as an estimate of the error covariances. There is also empirical evidence to suggest that these procedures are more robust when distributional assumptions are violated. A number of other estimation techniques (Two Stage Least Squares) have been proposed for Seemingly Unrelated Regression models, which form a subclass of MAGs

7.1 *OLS* ESTIMATION PROCEDURE FOR *MAGS/DAGS*

Let $x_1 \ldots x_p$ denote p variables with n observations on each. Given a MAG (or as a special case a DAG) model for the p variables we consider the class of multivariate normal distributions $N_p(\mu, \Sigma)$, such that the conditional independencies corresponding to the d-separations in the model are satisfied. This amounts to imposing certain constraints on the elements of Σ only,

with the mean vector playing no role. More specifically a DAG or a MAG corresponds to a class of multinormal distributions with μ unconstrained and Σ varying in a subset $\mathcal{B}$ of the class of $p \times p$ positive definite matrices. Let $X_{(p\times n)}$ denote the data matrix on the p variables and n observations. The centered sum of squares and product matrix for the data is then given by $S = X(I_n \perp \frac{\mathbf{1}_{n\times n}}{n})X^{\mathrm{T}}$. If we now consider the problem of maximising the log of the joint likelihood over all μ and Σ varying in $\mathcal{B}$ (this being precisely the subclass for which the constraints imposed by the d-separation relations in the graph hold) it is easy to see that the maximisation can be done by setting μ to be equal to the sample mean vector $\bar{X}$ (this is the maximum likelihood estimate for μ) and then maximizing the resulting expression over Σ. More specifcally we need to find:

$$\sup_{\Sigma\in\mathcal{B}} (\perp\frac{np}{2}\log(2\pi) \perp \frac{n}{2}\log(|\Sigma|) \perp \frac{1}{2}\mathrm{tr}\Sigma^{-1}S)$$

If we now think of having been originally provided with the centered data (the data centered around the sample mean vector) then the expression enclosed in brackets above is precisely the log of the joint likehihood for the centered data under a multivariate $N_p(0,\Sigma)$ density. Thus it suffices to restrict to the mean 0 case for purposes of maximization and look at parametrizations of $N_p(0,\Sigma)$ distributions with $\Sigma \in \mathcal{B}$

Given a MAG or DAG we can parametrize the class of $N_p(0,\Sigma)$ distributions corresponding to the d-separations in the MAG(DAG)in the following way.

Let $(\epsilon_1, \epsilon_2 \ldots, \epsilon_p)$ be random variables with $\epsilon \sim N_p(0,\Lambda)$. The double headed arrows are then given the interpretation of correlations between the ϵ's for the corresponding vertices. Thus, for a DAG, Λ has only diagonal non-zero entries(since double-headed arrows are absent). For a MAG, Λ has nonzero off-diagonal entries in its (i,j)'th position corresponding to the covariance between ϵ_i and ϵ_j whenever there is a double headed arrow between x_i and x_j. Let further $pa(x_i)$ denote the parents of variable x_i in the MAG. Then write x_i as

$$x_i = \sum_{x_j\in pa(x_i)} \alpha_j x_j + \epsilon_i.$$

Suppose further that $x_1 \ldots x_n$ are named such that no vertex precedes its parents, i.e. $\mathrm{pa}(x_i) \subseteq \{x_1 \ldots x_{i-1}\}$, $\forall i$. Then we can write:

$$V = \alpha_{p\times p} V_{p\times p} + \epsilon_{p\times 1},$$

where V is $(x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, α is a strictly lower triangular matrix with the i'th row of α having non-zero entires only in those positions that correspond to the parents of x_i, these entries being precisely the "regression coefficients" of x_i on its parents and $\epsilon \sim N(0,\Lambda)$. Thus

$$\epsilon = (I \perp \alpha)V,$$

or

$$V = (I \perp \alpha)^{-1}\epsilon.$$

Note that the inverse exists because $(I \perp \alpha)$ is lower triangular with 1's on the diagonal.
Therefore $V \sim N_p(0, (I \perp \alpha)^{-1}\Lambda((I \perp \alpha)^{-1})^{\mathrm{T}})$ and

$$\Sigma = (I \perp \alpha)^{-1}\Lambda((I \perp \alpha)^{-1})^{\mathrm{T}}.$$

The number of free parameters which is the dimension of the model is clearly the number of variables p + number of edges in the graph. Each edge corresponds to a regression coefficient or a covariance between errors and there are the p variances of the errors too and so we need to add p to the number of edges. Estiamtion techniques come up with estimates of α and Λ and use these to estimate Σ.

Suppose x_i is regressed on its parents $pa(x_i)$. Let $S_{pa(x_i)\cup\{x_i\}}$ be the centered sums of squares matrix for $pa(x_i) \cup \{x_i\}$, which we partition as follows:

$$S_{pa(x_i)\cup\{x_i\}} = \begin{bmatrix} S_{pa(x_i)} & \mathbf{r}_{x_i} \\ \mathbf{r}_{x_i}^{\mathrm{T}} & s_{x_i} \end{bmatrix}$$

where $S_{pa(x_i)}$ is the sums of squares matrix for the parents of i, $\mathbf{r}_{x_i}$ is an $l \times 1$ vector, and s_{x_i} is the sum of squares for x_i. To get the estimates of α_i and σ_i^2 the variable x_i is regressed on its parents, i.e.

$$\hat{\alpha} = S_{\mathrm{pa}(x_i)}^{-1}\mathbf{r}_{x_i}$$

and

$$\hat{\sigma}_i^2 = \frac{1}{n}(s_{x_i} \perp \mathbf{r}_{x_i}^{\mathrm{T}} S_{pa(x_i)}^{-1}\mathbf{r}_{x_i}).$$

In case of a MAG where there is a double headed arrow, the off-diagonal elements of the covariance matrix Λ can be estimated as

$$\hat{\sigma}_{ij} = \frac{1}{N}\sum_{t=1}^{n}(x_{it} \perp \hat{x}(\mathrm{pa}(x_i))_t)(x_{jt} \perp \hat{x}(\mathrm{pa}(x_j))_t)$$

where $\hat{x}(\mathrm{pa}(x_i)_t)$ is the fitted value for the t-th observation, given by regressing x_i on $\mathrm{pa}(x_i)$. Thus σ_{ij} is estimated by the covariance between the residuals after regressing x_i on its parents, and the residuals after regressing x_j on its parents.

7.2 PRELIMINARY EMPIRICAL RESULTS

It is important to note that while the OLS estimates are MLEs for DAG models, this is no longer true for MAGs in general. Consequently, the likelihood evaluated at the estimated OLS values may give different values for Markov equivalent MAGs. If this effect is large enough it could prove a major disadvantage to using OLS based estimates in a score when performing a specification search. This is particularly true of a search across equivalence classes.

We are currently investigating the difference between scores based on evaluating the likelihood at the MLE found via an iterative numerical algorithm, vs. evaluating the likelihood at the OLS estimates. Figure 7 shows LRT statistic evaluated at the MLE vs. the LRT statistic evaluated at the OLS estimates, for several different MAG models on the Cochran data set.

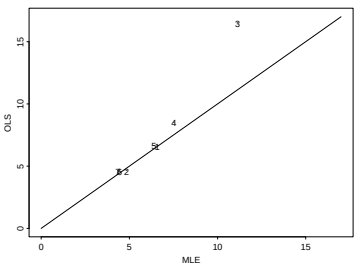

Figure 7: the LRT evaluated at the MLE vs. the LRT evaluated at the OLS estimates for

These results are preliminary, but suggest reasonable agreement between the scores. A larger simulation study is underway, but the results are incomplete.

8 *SUR* MODELS

Seemingly Unrelated Regression Models arise when we measure more than one response variable on a group of individuals. So suppose that we have n individuals, on each of whom m measurements are made and let $Y_1, Y_2, \ldots, Y_m$ (each an $n \times 1$ vector) denote the vectors of measurements. Also suppose that corresponding to each Y_i we have k_i predictor variables and let $X_{i_{(n \times k_i)}}$ denote the matrix of predictors. We also denote the sum of the k_i's by K. We can now formulate m regression models:
For $i = 1, 2, \ldots, m$

$$Y_i = X_i \beta_i + U_i$$

with $\mathrm{E}[U_i] = 0$ for all i and j $\mathrm{E}[U_i U_j^{\mathrm{T}}] = \sigma_{ij} I_{n \times n}$, where $\Sigma = ((\sigma_{ij}))$ is a positive definite matrix. This amounts to stipulating that if $\tilde{U}_k = (u_{k1}, u_{k2}, \ldots, u_{km})^{\mathrm{T}}$ is the vector of errors for measurements on the k'th individual then $E[\tilde{U}_k] = 0$ and $\mathrm{Cov}(\tilde{U}_k) = \Sigma$. Further $\tilde{U}_k$ and $\tilde{U}_l$ are uncorrelated if $k \neq l$. Under the normality assumption on the errors we have $\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_n$ are i.i.d. $N(0, \Sigma)$. Thus errors on the same individual corresponding to different measurements are allowed to be correlated. The expression 'seemingly unrelated regression model' because the m equations are related to one another even though superficially they may not seem to be so.
One standard method of estimating the regression coefficients is to use the standard OLS coefficients;thus the estimate of β_i is $\hat{\beta}_i = (X_i^{\mathrm{T}} X_i)^{-1} X_i^{\mathrm{T}} Y_i$. In this case the β_i's are estimated independently of each other. The residuals from the the i-th regression are $\hat{U}_i = Y_i \perp X_i \hat{\beta}_i$ and these can be used to estimate the parameters σ_{ij}. More specifically $\hat{\sigma}_{ij} = n^{-1} \hat{U}_i^{\mathrm{T}} \hat{U}_j$. However since we have a structural dependence of these m equations through the elements of Σ, it is indeed possible to construct more efficient estimates of the regression coefficients by exploiting this dependence. The m regression equations can be viewed as one giant regression equation:

$$Y_{nm \times 1} = X_{nm \times K} \beta_{K \times 1} + U_{nm \times 1},$$

where

$$Y = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, \ldots, Y_m^{\mathrm{T}})^{\mathrm{T}}$$
$$U = (U_1^{\mathrm{T}}, U_2^{\mathrm{T}}, \ldots, U_m^{\mathrm{T}})^{\mathrm{T}}$$
$$\beta = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}, \ldots, \beta_m^{\mathrm{T}})^{\mathrm{T}}$$
$$X_{nm \times K} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & X_m \end{bmatrix}$$

this is the giant matrix of predictors such that if it is partitioned by grouping the rows in stacks of size n and grouping the columns in stacks of size $k_1, k_2, \ldots, k_m$ respectively, then the diagonal entries of the partitioned matrix (in which the entries themselves are submatrices) are $X_1, X_2, \ldots, X_n$ respectively and the off-diagonal entries are 0. We note also that $E[U] = 0$ and $\mathrm{Cov}[U] = \Sigma \otimes I_n$. One then immediately obtains the generalized least squares estimate of β by transforming the above regression equation to one with homoscedastic errors. The GLS estimate is:

$$\hat{\beta}_{GLS} = (X^{\mathrm{T}} (\Sigma^{-1} \otimes I_n) X)^{-1} X^{\mathrm{T}} (\Sigma^{-1} \otimes I_n) Y.$$

In general Σ is unknown but one usually plugs in a consistent estimator of Σ into the above equation, say $\hat{\Sigma} = ((\hat{\sigma}_{ij}))$ where $\hat{\sigma}_{ij}$ is obtained using the residuals from OLS regression as explained previously. Both

$\hat{\beta}_{OLS}$ and $\hat{\beta}_{GLS}$ are unbiased and in particular under the normality assumption the estimato obtained by plugging in $\hat{\Sigma}$ for Σ in $\hat{\beta}_{GLS}$ is unbiased too. Under the normality assumption with no restrictions on Σ one can frame the likelihood equations as follows:

$$X^{\mathrm{T}}(\Sigma^{-1} \otimes I_n)X\beta = X^{\mathrm{T}}(\Sigma^{-1} \otimes I_n)Y$$

$$\Sigma = n^{-1}((u_{ij})); \;\; u_{ij} = [(Y_i \perp X_i\beta_i)^{\mathrm{T}}(Y_j \perp X_j\beta_j)]$$

Even though analytical closed form expressions for the MLE's do not exist the following iterative procedure gives satisfactory results:
$\hat{\beta}_w = (X^{\mathrm{T}}(S_{(w)}^{-1} \otimes I_n)X)^{-1}X^{\mathrm{T}}(S_{(w)}^{-1} \otimes I_n)Y; w = 1, 2, \ldots$
$\hat{U}_{i(w)} = Y_i \perp X_i\hat{\beta}_{i(w-1)}; w = 2, 3, \ldots$
$S_{ij\,(w)} = n^{-1}[\hat{U}_{i(w)}^{\mathrm{T}}\hat{U}_{j(w)}]; w = 2, 3, \ldots$, where one iterates till convergence. The initial estimate of Σ, i.e. $S_{(1)}$ can be obtained from the residuals of OLS regression.

9 APPLYING *SUR* METHODS TO GAUSSIAN *MAG* VIA

9.1 KEY ISSUES

The SUR techniques described earlier can be adapted to estimation of parameters in a subclass of Gaussian MAGs defined below.

At first sight it might appear that a MAG model on m vertices is reducible to a set of m seemingly unrelated regression equations, with the parents of each vertex acting as the predictors in the corresponding regression equation.

However in a MAG some responses may act as predictors for other responses whereas this may not occur in a SUR model (where the predictors are usually considered fixed). A second difference is that in a MAG the error terms in two equations may be specified to be uncorrelated, while in a SUR model the covariance matrix for the disturbances is unrestricted.

To apply the SUR model correctly to a subset $\mathbf{B}$ of vertices, we also require that there be no double headed arrow between any vertex in $\mathbf{B}$ and any vertex in the set of strict ancestors of $\mathbf{B}$. (Where the *strict ancestral set* fo $\mathbf{B}$ is defined as $\bigcup_{x \in \mathbf{B}}(\mathrm{an}(b) \setminus \{b\}$.) This is necessary because while using the SUR model with the vertices in $\mathbf{B}$ as response we are modelling the conditional distribution of the response variables given their parents and so are estimating the conditional covariance matrix between the error variables given the parents. But we need to estimate the unconditional error covariance matrix and this requires the independence of the errors corresponding to the response and the parents of $\mathbf{B}$. This is no longer guaranteed if some error corresponding to a vertex in the strict ancestral set of $\mathbf{B}$ is correlated with some error corresponding to a vertex in $\mathbf{B}$ (since every parent of $\mathbf{B}$ can be written as a linear combination of errors from the strict ancestral set). We now characterise classes of MAGS to which SUR methods can be applied.

9.2 FORMULATION

Consider a MAG model on m vertices. Suppose that the vertex set admits a partitioning into k subsets $\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_k}$ such that the following properties are satisfied:

(a) In each block none of the vertices is an ancestor of any other.

(b) For any given block $\mathbf{V}$, denote its connected components by $\mathbf{C_1}, \ldots \mathbf{C_l}$. We assume that there is no double headed arrow between $\mathbf{C_i}$ and the strict ancestral set of $\mathbf{C_i}$, for all i.

We can then employ the SUR model as follows:

For any given block $\mathbf{V}$ look at a connected component $\mathbf{C}$. Since none of the vertices in $\mathbf{V}$ is an ancestor of any other, the only connections between vertices in $\mathbf{C}$ are double-headed arrows and these correspond to possibly non-zero entries in the covariance matrix for the errors associated with these vertices. Estimate the regression coefficients and the elements of the covariance matrix for the errors using SUR. If C is a clique, then the error covariance matrix is unconstrained and the iterative techniques for obtaining the MLE's of the β's and σ's as outlined in section 8 can be resorted to directly.

Otherwise there are constraints on the error covariance matrix and we may extend the iterative estimation technique in the following way:

Compute the estimate of the error covariance σ_{ij} at each stage in the same way as outlined for the iterative MLE computation if there is a double-headed arrow between vertices i and j. Otherwise force $\sigma_{i,j}$ to be equal to zero. Repeat process till convergence. We note however that there is no guarantee that the process converges to the MLE in this situation. But the estimates obtained in the above manner ought to be consistent.

In this way we obtain final estimates of the edge coefficients and error covariances for each connected component within a block. There now remains the question of estimating the covariances between errors that belong to different blocks and this is done by computing the sample covariance between the residuals for the corresponding vertices (where the final estimates of the betas are used to compute the residuals).

9.3 TWO SPECIAL CASES

(1) One way of stratifying vertices into subsets so that within each subset there is no directed edge or path, is to sort them by their generation. Define a vertex to belong to *generation i* if the length of the longest directed path entering that vertex in the MAG is i. It is then easy to see that no vertex in the i-th generation can be a parent of any other. One can then check for condition (b) and provided it is satisfied estimation can be carried out. In particular, condition (b) is satisfied if there are no double-headed arrows between vertices of different generations.

(2) Define an equivalence relation on the set of vertices in the following way: a is in the same equivalence class as b if a can be reached from b by a path consisting entirely of double-headed arrows. Let $\mathbf{V_1}, \mathbf{V_2}, \ldots \mathbf{V_k}$ denote the equivalence classes under this relation. Note that under this stratification each $\mathbf{V_i}$ is connected. Note also that if condition (a) is satisfied in this case there is no further need to check condition (b). It will be satisfied automatically (this is a direct outcome of the construction of these blocks). One can then use SUR techniques as described.

Acknowledgements

A grant from the National Science Foundation supported this work (DMS 9704573), and from a Royalty Research Fund Grant from the University of Washington. The first author was a Rosenbaum fellow at the Isaac Newton Institute, Cambridge, England from July-December 1997, where part of this work was undertaken. Peter Spirtes and Nanny Wermuth made many helpful suggestions. Michael Perlman and Lang Wu provided helpful input on SUR models.

References

Cochran, W.G. (1938). The omission or addition of an independent variate in multiple linear regression. *JRSS Supplement*, **5**, pp.171-176.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**, pp. 309-347.

Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman and Hall.

Chickering, D. and Geiger, D. and Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. *Preliminary papers of the fifth international workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, pp. 112-128.

Chickering, D. (1995) Learning equivalence classes of Bayesian network structures. *UAI* 11 (P. Besnard and S. Hanks, eds.), 150-157. Morgan Kaufmann: San Francisco, CA.

Dempster, A.P. (1972). Covariance Selection. *Biometrics*, **28**, pp. 157-175.

Geiger, D. and Meek, C. (1998) Graphical Models and Exponential Families. *UAI 14*, Morgan Kaufmann: San Francisco, CA.

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

Pearl, J. and Verma, T., A Theory of Inferred Causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference,* Morgan Kaufmann, pp. 441-452, April 1991.

Spirtes, P. and Cooper, G. (1999) An experiment in causal discovery using a pneumonia database. This volume.

Spirtes, P., Glymour, C., and Scheines, R., (1993) *Causation, Prediction, and Search,* Lecture Notes in Statistics **81**, Springer-Verlag, New York.

Spirtes, P., and Meek, C. (1995). Learning Bayesian Networks with Discrete Variables from Data. *In Proceedings of The First International Conference on Knowledge Discovery and Data Mining*, U. Fayyad and R. Uthurusamy, ed.s, AAAI Press, pp. 294-299.

Spirtes, P., and Richardson, T. (1997). A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Preliminary Papers, 6th International Workshop on Artificial Intelligence and Statistics* (D. Madigan and P. Smyth, eds.)

Spirtes, P., Richardson, T., and Meek, C. (1997). Heuristic Greedy Search Algorithms for Latent Variable Models. In Preliminary Papers, 6th International Workshop on Artificial Intelligence and Statistics (D. Madigan and P. Smyth, eds.).

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *UAI 6* pp.220-227. Morgan Kaufmann: San Francisco.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, NJ.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161-215.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Amer. Statist. Assoc.*, **57**, 348-368.