The Dimensionality of Mixed Ancestral Graphs

by Peter Spirtes, Thomas Richardson and Chris Meek

First we will introduce some graph terminology. The concepts defined here are illustrated in Figure 1. A graph consists of two parts, a set of vertices **V** and a set of edges **E**. Each edge in **E** is between two distinct vertices in **V**. There are two kinds of edges in **E**, directed edges A → B or A ← B, and double-headed edges A ↔ B; in either case A and B are **endpoints** of the edge; further, A and B are said to be **adjacent**. In Figure 1 the set of vertices is {A,B,C,D,E} and the set of edges is {A ↔ B, B → C, C → D, E → D}. For a directed edge A → B, A is the **tail** of the edge and B is the **head** of the edge, A is a **parent** of B, and B is a **child** of A.

An **undirected path** U between X_1 and X_n is a sequence of edges $<E_1,...,E_m>$ such that one endpoint of E_1 is X_1, one endpoint of E_m is X_n, and for each pair of consecutive edges E_i, E_{i+1} in the sequence, $E_i \neq E_{i+1}$, and one endpoint of E_i equals one endpoint of E_{i+1}. In Figure 1, A ↔ B → C ← D is an example of an undirected path between A and D. A **directed path** P between X_1 and X_n is a sequence of directed edges $<E_1,...,E_m>$ such that the tail of E_1 is X_1, the head of E_m is X_n, and for each pair of edges E_i, E_{i+1} adjacent in the sequence, $E_i \neq E_{i+1}$, and the head of E_i is the tail of E_{i+1}. For example, B → C → D is a directed path. A **vertex occurs on a path** if it is an endpoint of one of the edges in the path. The set of vertices on A ↔ B → C → D is {A, B, C, D}. A path is **acyclic** if no vertex occurs more than once on the path. The following is a list of all the acyclic directed paths in Figure 1: B → C, C → D, E → D, B → C → D.

A graph is a **directed graph** if it contains no double-headed edges.A graph is a **directed acyclic graph** (DAG) if it contains no double-headed edges, and no directed cycles.

A vertex A is an **ancestor** of B (and B is a **descendant** of A) if and only if either there is a directed path from A to B or A = B. Thus the ancestor relation is the transitive, reflexive closure of the parent relation. The following table lists the child, parent, descendant and ancestor relations in Figure 1.

Vertex	Children	Parents	Descendants	Ancestors
A	$\varnothing$	$\varnothing$	{A}	{A}
B	{C}	$\varnothing$	{B,C,D}	{B}
C	{D}	{B}	{C,D}	{B,C }
D	$\varnothing$	{C,E}	{D}	{B,C,D,E}
E	{D}	$\varnothing$	{D,E}	{E}

A vertex X is a **collider** on undirected path U if and only if U contains a subpath $Y \leftrightarrow X \leftrightarrow Z$, or $Y \rightarrow X \leftrightarrow Z$, or $Y \rightarrow X \leftarrow Z$, or $Y \leftrightarrow X \leftarrow Z$; otherwise if X is on U it is a **non-collider** on U. For example, D is a collider on $C \rightarrow D \leftarrow E$ and C is a non-collider on $B \rightarrow C \rightarrow D$. X is an **ancestor of a set** of vertices **Z** if X is an ancestor of some member of **Z**.

For disjoint sets of vertices, **X**, **Y**, and **Z**, **X** is **d-connected** to **Y** given **Z** if and only if there is an acyclic undirected path U between some member X of **X**, and some member Y of **Y**, such that every collider on U is an ancestor of **Z**, and every non-collider on U is not in **Z**. For disjoint sets of vertices, **X**, **Y**, and **Z**, **X** is **d-separated** from **Y** given **Z** if and only if **X** is not d-connected to **Y** given **Z**.

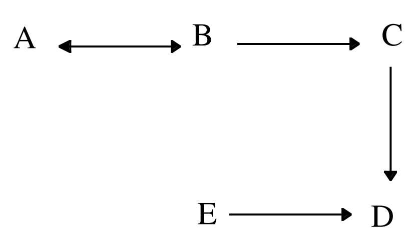

Figure 1

For example, the path $A \leftrightarrow B \rightarrow C$ d-connects A and C given $\varnothing$; it also d-connects A and C given {D}, {E}, or {D,E}. $E \rightarrow D \leftarrow C$ d-connects E and C given {D}, given {D,B}, {D,A}, or {D,A,B}. The following is a list of all the pairwise d-separation relations in Figure 1 (where each pair is followed by a list of all of the sets that d-separate them):

{A} and {C} are d-separated given: {B}, {B,D}, {B,E}, {B,D,E}

{A} and {D} are d-separated given: {B}, {C}, {B,C}, {B,E}, {C,E}, {B,C,E}

{A} and {E} are d-separated given: ∅, {B}, {C}, {B,C}, {B,D}, {C,D}, {B,C,D}

{B} and {E} are d-separated given: ∅, {A}, {C}, {A,C}, {C,D}, {A,C,D}.

In a graph G, with a set of vertices **V** containing **O**, if A and B are in **O**, then there is an **inducing path** between A and B given **O** if and only if there is a path U between A and B such that every member of **O** that is on U is a collider, and every collider is an ancestor of A or B. (If **V** = **O**, we will simply say that there is an inducing path between A and B.) It has been shown in Verma and Pearl(1990) that in a DAG G, A and B are d-separated given some subset of **O**\{A,B} if and only if there is no inducing path between A and B given **O**.

A **MAG** (or **mixed ancestral graph**) is a graph with two kinds of edges: directed edges (e.g. A → B), and bi-directed edges, (e.g. C ↔ D). The MAG that represents a DAG G (also called MAG(G,**O**) with a set of observed variables **O**) can be constructed in the following way:

- Place the edge A → B in MAG(G,**O**) if and only if A is an ancestor of B in G, and there is an inducing path between A and B given **O** in G.
- Place the edge A ↔ B in MAG(G,**O**) if and only if A is not an ancestor of B in G, B is not an ancestor of A in G, and there is an inducing path between A and B given **O** in G.

Some examples of MAGs are shown in Figure 2, where **O** = {A,B,C,D}. (In cases where the distinction between latent variables and measured variables is important, we enclose latent variables in ovals.)

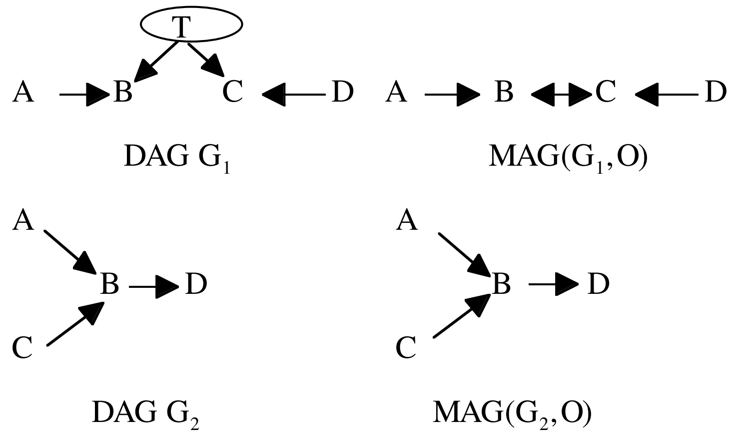

Figure 2

It has been shown in Spirtes et al. (1993) that in a MAG G, A and B are d-separated given some subset of {A,B} if and only if there is no inducing path between A and B in G. (In Spirtes et al. 1993 the proof is for inducing path graphs, but the proofs carry over unchanged for MAGs.)

Thus a MAG M may be considered to represent any DAG G such that M = MAG(G,**O**). MAG(G,**O**) represents the following features of a DAG G with latent variables:

- the ancestor relations among the members of **O** in G;
- the d-separation relations among the members of **O** in G.

Although we will not prove it in this paper, MAGs have the following useful features:

- DAG G_1 in Figure 2 is an example of a DAG such that as the sample size increases without limit, the difference between the BIC of MAG(G_1,**O**) and the BIC of any DAG G' that contains only variable in **O** increases without limit almost surely. Hence in some cases a maximum likelihood estimate of the MAG parameters is a better estimator of some of the population parameters than the maximum likelihood estimate of any DAG parameters.
- In the large sample limit, for multi-variate normal, any (possibly latent variable) DAG with a maximum BIC score is represented by the MAG with the highest BIC score among all MAGs.

In general, each MAG represents many different latent variable models. A MAG can thus also be considered a representation of a set of conditional independence relations among variables in **O** (which in some cases cannot be represented by any DAG containing just variables in **O**.) A MAG imposes no restrictions on the set of distributions it represents other than the conditional independence relations that it entails. (The class of MAGs is neither a subset nor a superset of other generalizations of DAGs such as chain graphs, cyclic directed graphs, or cyclic chain graphs.)

It is not the case that an arbitrary graph is a MAG, i.e. there may be no DAG G such that M = MAG(G,**O**). The following theorem states necessary and sufficient conditions for M to be a MAG. Let the **canonical graph** G(M) for a MAG M be constructed in the following way: If **O** is the set of vertices in M, then the set of vertices in G(M) is $\mathbf{O} \cup \mathbf{T}$, where $T_{i,j} \in \mathbf{T}$ if there is an edge $X_i \leftrightarrow X_j$ in M, for X_i, X_j in **O**, there are edges $X_i \leftarrow T_{i,j} \rightarrow X_j$ in G(M) if and only if there is an edge $X_i \leftrightarrow X_j$ in M, and an edge $X_i \rightarrow X_j$ in G(M) only if there is an edge $X_i \rightarrow X_j$ in M.

Lemma 1: If M is a MAG, then G(M) has the same ancestor relations among members of **O** as M does.

Proof. By the algorithm for constructing G(M), the set of directed edges in G(M) is a superset of directed edges in M. Hence if there is a directed path in M, there is a corresponding directed path in G(M).

Suppose there is a directed path P in G(M) between X_i and X_j in **O**. P does not contain any member of **T**, because by the algorithm for constructing G(M), every edge containing a member of $T_{a,b}$ of **T** is out of $T_{a,b}$. Hence every edge on P is a directed edge between members of **O**. By the algorithm for constructing G(M), there is a corresponding directed edge in M. Hence there is a directed path between X_i and X_j in G(M). $\therefore$

Note that G(M) is acyclic because there is no directed cycles from a member of **O** to itself in M, and hence no directed cycles from a member of **O** to itself in G(M). Also there is no directed cycle from a member of **T** to itself in G(M), because there are no edges into a member of **T** in G(M).

Lemma 2: If M = MAG(G) has vertices **O**, then M and G have the same ancestor relations among members of **O**.

Proof. Suppose there is a directed path P from X_i to X_j in M. Then by the algorithm for constructing MAGs, each vertex along P is an ancestor of its sucessor on the path in G. Hence X_i is an an ancestor of X_j in G.

Suppose there is a directed path P from X_i to X_j in G. Then if X_a and X_b are on P, and there is no other member of **O** between X_a and X_b on P, there is an inducing path relative to **O** in G between X_a and X_b. It follows that there is an edge $X_a \rightarrow X_b$ in M. The concatenation of these edges in M is a directed path from X_i to X_j in M. $\therefore$

Theorem 1: A graph M is a MAG if and only if:

1. If there is an inducing path between X_i and X_j in M, then X_i and X_j are adjacent in M.
2. If there is an edge $X_i \rightarrow X_j$ in M, then X_j is not an ancestor of X_i in M, but X_i is an ancestor of X_j in M.
3. If there is an edge $X_i \leftrightarrow X_j$ in M, then X_j is not an ancestor of X_i and X_i is not an ancestor of X_j in M.

Proof. Suppose that M satisfies 1), 2), and 3). We will show that there is some graph G(M) such that MAG(G(M)) = M.

First we will show that if there is an inducing path U between X_i and X_j in G(M), then X_i and X_j are adjacent in M. Let U' be the path in M corresponding to U (that is if $X_i \rightarrow X_j$ occurs on U in G(M), $X_i \rightarrow X_j$ occurs on U' in M, and if $X_i \leftarrow T_{i,j} \rightarrow X_j$ occurs on U in G(M) then $X_i \leftrightarrow X_j$ occurs on U' in M.) By definition, every member of **O** on U in G(M) is a collider on U. From the algorithm for constructing G(M), it follows that every member of **O** on U' in M is a collider. Every collider on U in G(M) is an ancestor of X_i or X_j in G(M). Because by Lemma 1, G(M) and M have the same ancestor relations, every collider on U' in M is an ancestor of X_i or X_j in M. By hypothesis, if there is an inducing path between X_i and X_j in M, X_i and X_j are adjacent in M.

Next we will show that if X_i is an ancestor of X_j in G(M), and X_i and X_j are adjacent in M, then the edge between X_i and X_j is oriented as $X_i \rightarrow X_j$ in M. Because G(M) is acyclic, X_j is not an ancestor of X_i in G(M). By Lemma 1, G(M) and M have the same ancestor relations, so X_i is an ancestor of X_j in M. By 2) and 3), the edge between X_i and X_j is oriented as $X_i \rightarrow X_j$ in M.

Finally we will show that if X_i is not an ancestor of X_j in G(M), X_j is not an ancestor of X_i in G(M),and X_i and X_j are adjacent in M, then the edge between X_i and X_j is oriented as $X_i \leftrightarrow X_j$ in M. By Lemma 1, G(M) and M have the same ancestor relations among members of **O**, so X_i is not an ancestor of X_j in G(M) and X_j is not an ancestor of X_i in M. By 3), the edge between X_i and X_j is oriented as $X_i \rightarrow X_j$ in M.

It follows that MAG(G(M)) = M.

Next we will show that conversely, if M is a MAG, then 1), 2), and 3) hold. Since M is a MAG, there is some graph G such that M = MAG(G).

First we will show 1). Suppose on the contrary that there is an inducing path between X_i and X_j in M, but X_i and X_j are not adjacent in M. Because there is an inducing path between X_i and X_j in M, X_i and X_j are d-connected given every subset of $\mathbf{O}\backslash\{X_i,X_j\}$ in M. In Spirtes and Richardson(1996) it was shown that if X_i and X_j are d-connected given **Z** in MAG(G), then X_i and X_j are d-connected given **Z** in G. It follows that X_i and X_j are d-connected given every subset of $\mathbf{O}\backslash\{X_i,X_j\}$ in G. By Verma and Pearl(1990), there is an inducing path between X_i and X_j relative to **O** in G. By the method of construction of MAGs, X_i and X_j are adjacent in M.

Next we show 2). Suppose that there is an edge $X_i \rightarrow X_j$ in M, but contrary to the hypothesis, X_j is an ancestor of X_i, or X_i not an ancestor of X_j in M. By the method of construction of MAGs, in G, X_i is ancestor of X_j, and X_j is not ancestor of X_i. By Lemma 2, M has the same ancestor relations as G. This is a contradiction.

Finally, we show 3). Suppose there is an edge $X_i \leftrightarrow X_j$ in M, but contrary to the hypothesis that X_j is an ancestor of X_i or X_i is an ancestor of X_j in M. In G, by the algorithm for construction MAGs, X_i is not ancestor of X_j and X_j not ancestor of X_i. By Lemma 2, M has the same ancestor relations as G. This is a contradiction. $\therefore$

A **linear parameterization of a MAG G** is a linear structural equation model with correlated errors parameterized in the following way:

- Each variable A in the graph is a linear function of its parents in the graph, and a unique error term, ε_A.
- Two error terms ε_A.and ε_B have a non-zero correlation only if there is an edge A $\leftrightarrow$ B in the graph.

For notational convenience, we will assume that the variables in a MAG G are $X_1,\ldots,X_n$, where X_i is not an ancestor of X_j in G if i > j. We will refer to the error term of X_i as ε_i, rather than ε_{X_i}. A MAG M **linearly entails** that **X** is independent of **Y** given **Z** if and only if in every linear parameterization of M, **X** is independent of **Y** given **Z**. The following theorems are proved in Spirtes and Richardson(1997).

Theorem 2: A MAG G linearly entails X_i is independent of X_j given **Z** if and only if X_i is d-separated from X_j given **Z** in G.

Theorem 3: A MAG G linearly entails $\text{cov}(X_i,X_j|\mathbf{Z}) = 0$ if and only if X_i is d-separated from X_j given **Z** in G.

A **complete MAG** is a MAG in which every pair of variables is adjacent.

Lemma 3: If G is a MAG, there is a complete MAG G_C, such that G is a subgraph of G_C.

Proof. Suppose that G is a MAG. Form G_C in the following way: if X_i is an ancestor of X_j in G, add an edge $X_i \rightarrow X_j$ in G_C, and if X_i is not an ancestor of X_j and X_j is not an ancestor of X_i, then add an edge $X_i \leftrightarrow X_j$ to G_C. Every pair of vertices in G_C is adjacent. Note that G and G_C have the same ancestor relations. By the method of construction of G_C, if there is an edge $X_i *\!\!\rightarrow X_j$ in G_C, then X_j is not an ancestor of X_i. Because every pair of vertices in G_C is adjacent, it is trivially true that if there is an inducing path between a pair of vertices, then there is an edge between that pair of vertices. $\therefore$

Lemma 4: In a MAG G with an edge $X_i \leftrightarrow X_j$, if $\theta(0)$ is a parameterization of G in which $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = 0$, and $\theta(c)$ is a parameterization of G identical to $\theta(0)$ except $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = c$, then

$$\mathrm{cov}_{\theta(c)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = \mathrm{cov}_{\theta(0)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) + c.$$

Proof. First we will show that if X_k is in $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}$ then $\mathrm{var}_{\theta(0)}(X_k) = \mathrm{var}_{\theta(c)}(X_k)$. Note that in $G(\theta(0))$ and $G(\theta(c))$, the coefficients of the reduced form of each variable is exactly the same in each of the parameterizations, because the structual equations in each parameterization are identical. X_k is a function of (a possibly proper subset of) the error terms in $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}$. Hence

$$X_k = \sum_{r\le k} a_{kr}\varepsilon_r$$

$$\mathrm{var}_{q(0)}(X_k) = \sum_{r\le k} a_{kr}^2 \mathrm{var}_{q(0)}(\varepsilon_r) + 2\sum\sum_{r<s\le k} a_{kr}a_{ks}\mathrm{cov}_{q(0)}(\varepsilon_r,\varepsilon_s)$$

In the formula for $\mathrm{var}_{\theta(0)}(X_k)$, r and s are less than or equal to k, and k is less than i and j by hypothesis. Hence, the formula is identical for $\mathrm{var}_{\theta(c)}(X_k)$, and hence $\mathrm{var}_{\theta(0)}(X_k) = \mathrm{var}_{\theta(c)}(X_k)$.

Next we will show that if distinct vertices X_k and X_l are in $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)$, but $\{X_k,X_l\}$ <> $\{X_i,X_j\}$, then $\mathrm{cov}_{\theta(0)}(X_k,X_l) = \mathrm{cov}_{\theta(c)}(X_k,X_l)$. Suppose X_k and X_l are in $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)$, but $\{X_k,X_l\}$ <> $\{X_i,X_j\}$. Also, suppose without loss of generality that X_l is not equal to X_i or to X_j and hence X_l is in $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}$. X_i is not an ancestor of any member of $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i\}$; if it is an ancestor of $\mathbf{An}(X_i)\backslash\{X_i\}$ then there is a directed cycle in G, and if it is an ancestor of $\mathbf{An}(X_j)\backslash\{X_i\}$ then X_i is an ancestor of X_j, even though there is an edge $X_i \leftrightarrow X_j$ in G. Similarly, X_j is not an ancestor of any member of $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_j\}$. It follows that neither X_i nor X_j is an ancestor of X_l.

For X_k and X_l,

$$X_k = \sum_{r\le k} a_{kr}\varepsilon_r, \quad X_l = \sum_{r\le l} b_{lr}\varepsilon_r$$

$$\mathrm{cov}_{q(0)}(X_k,X_l) = \sum_{r\le k} a_{kr}b_{lr}\mathrm{var}_{q(0)}(\varepsilon_r) + 2\sum\sum_{r<s\le \max(k,l)} (a_{kr}b_{ls} + a_{ks}b_{lr})\mathrm{cov}_{q(0)}(\varepsilon_r,\varepsilon_s)$$

The formula for $cov_{\theta(c)}(X_k,X_l)$ is exactly the same as the formula for $cov_{\theta(0)}(X_k,X_l)$ except for terms of the form $(a_{ki}b_{lj} + a_{kj}b_{li})cov\theta_{(c)}(\varepsilon_i,\varepsilon_j)$; all corresponding terms are zero in $cov_{\theta(0)}(X_k,X_l)$ because $cov\theta_{(0)}(\varepsilon_i,\varepsilon_j) = 0$ by definition. However, all such terms are also zero in the formula for $cov_{\theta(c)}(X_k,X_l)$, because neither X_i nor X_j is an ancestor of X_l, and henceb_{lj} and $b_{li} = 0$.

Next we will show that $cov_{\theta(0)}(X_i,X_j) = cov_{\theta(c)}(X_i,X_j) + c$. The formula for $cov_{\theta(0)}(X_i,X_j)$ is identical to the formula for $cov_{\theta(c)}(X_i,X_j)$, except that the term $a_{ii}b_{jj}cov\theta_{(c)}(\varepsilon_i,\varepsilon_j) = cov\theta_{(c)}(\varepsilon_i,\varepsilon_j) = c$, and the term $a_{ii}b_{jj}cov\theta_{(0)}(\varepsilon_i,\varepsilon_j) = cov\theta_{(0)}(\varepsilon_i,\varepsilon_j) = 0$. (By convention, the error terms are scaled so that $a_{ii} = b_{ii} = 1$.

$$\begin{gathered}
cov_{\theta(0)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = \\
cov_{\theta(0)}(X_i,X_j) - cov_{\theta(0)}(X_i\mathbf{,}\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) \times \\
var_{\theta(0)}^{-1}(\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) \times \\
cov_{\theta(0)}(X_j\mathbf{,}\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\})
\end{gathered}$$

The formula for $cov_{\theta(0)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\})$ is the same except everywhere $\theta(0)$ occurs it is replaced by $\theta(c)$. We have just shown that:

$$\begin{gathered}
cov_{\theta(0)}(X_i,\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = cov_{\theta(c)}(X_i,\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) \\
var_{\theta(0)}^{-1}(\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = var_{\theta(c)}^{-1}(\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) \\
cov_{\theta(0)}(X_j\mathbf{,}\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = cov_{\theta(c)}(X_j\mathbf{,}\ \mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) \\
cov_{\theta(0)}(X_i,X_j) = cov_{\theta(c)}(X_i,X_j) + c
\end{gathered}$$

Hence,

$$cov_{\theta(c)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = cov_{\theta(0)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) + c.$$

$\therefore$

Theorem 4: If G_C is a complete MAG over a set of variables $\mathbf{X}$, and Σ is a positive definite covariance matrix for $\mathbf{X}$, then there is a linear parameterization θ of G_C such that $\Sigma_{G_C(\theta)} = \Sigma$.

Proof. Let Σ be the covariance matrix for $\mathbf{X}$. An instantiation of a parameterization of G_C has the properties that each variable can be expressed as linear function of its parents and an error term, and that if $\text{cov}(\varepsilon_p,\varepsilon_q) \neq 0$ then $X_p \leftrightarrow X_q$ in G_C; we will now show that there is a parameterization of G_C that has covariance matrix Σ.

Note that since G_C is a complete ancestral graph $\textbf{Parents}(X_k) \subseteq \{X_j|j < k\}$, and further if $X_i \in \{X_j|j < k\}\backslash\textbf{Parents}(X_k)$ then $X_i \leftrightarrow X_j$ in G_C. We will abbreviate $\textbf{Parents}(X_k)$ by $\mathbf{P_k}$. Take each variable X_k in turn. Regress X_k on $\mathbf{P_k}$. Let

$$\hat{X}_k = \sum_{X_j \in \mathbf{P}_k} \alpha_{kj} X_j \text{ and}$$

be the linear predictor of X_k on $\mathbf{P_k}$ (where summation over an empty set is equal to zero) and the residuals

$$\varepsilon_k := X_k - \hat{X}_k$$

We will now show that the α_{kj} and the correlations between the residuals form a parameterization of the complete MAG G_C. First note that X_k is a linear function of its parents in G_C because

$$X_k = \hat{X}_k + \varepsilon_k = \sum_{X_j \in \mathbf{P_k}} \alpha_{kj} X_j + \varepsilon_k$$

Second, we will show that if $\text{Cov}(\varepsilon_p,\varepsilon_q) \neq 0$ then $X_p \leftrightarrow X_q$ in G_C. Suppose on the contrary that $\text{cov}(\varepsilon_p, \varepsilon_q) \neq 0$, but there is no double headed arrow $X_p \leftrightarrow X_q$ in G_C. We may suppose without loss of generality that $p < q$. Since there is no double headed arrow $X_p \leftrightarrow X_q$, and $p < q$ it follows that $X_p \rightarrow X_q$ in G_C. It then follows that $X_p \in \mathbf{P}_q$.

$$\text{cov}(\varepsilon_q,\varepsilon_p) = \text{cov}(\varepsilon_q, X_p - \sum_{X_j \in \mathbf{P_p}} X_j) = \text{cov}(\varepsilon_q, X_p) - \sum_{X_j \in \mathbf{P_p}} \text{cov}(\varepsilon_q, X_j)$$

We will now show that $\text{cov}(\varepsilon_q,\varepsilon_p) = 0$ by showing that $\text{cov}(\varepsilon_q,X_p) = 0$, and for all X_j in $\mathbf{P_p}$, $\text{cov}(\varepsilon_q,X_j) = 0$.

By construction, ε_q is uncorrelated with $X_p \in \mathbf{P}_q$, (since ε_q is the residual remaining after regressing X_q on $\mathbf{P}_q$), so $\text{cov}(\varepsilon_q,X_p) = 0$. If $X_j \in \mathbf{P}_p$, then $X_j \rightarrow X_p$ in G_C. Since $X_p \rightarrow X_q$ in G_C, it follows that X_j is an ancestor of X_q in G_C. As G_C is a complete ancestral graph it then follows that $X_j \rightarrow X_q$ in G_C, so $X_j \in \mathbf{P}_q$. Hence $\text{cov}(\varepsilon_q,X_j) = 0$, as claimed. It follows that $\text{cov}(\varepsilon_q,\varepsilon_p) = 0$.

Finally, positive definiteness of Σ ensures that each ε_k has positive variance; otherwise $X_k$would be a linear combination of previous X_i's and Σ would not be positive definite. $\therefore$

Lemma 5: In a MAG G, if $X_j \in \mathbf{An}(X_i)$, and there is no edge $X_j \to X_i$, then X_i is d-separated from X_j given $\mathbf{An}(X_i)\backslash\{X_i,X_j\}$.

Proof. Suppose, on the contrary that there is a path U that d-connects some member $X_j \in \mathbf{An}(X_i)$ to X_i given $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$. There are three cases: either there is an edge $X_k \to X_i$ on U, there is an edge $X_i \to X_k$ on U, or there is an edge $X_i \leftrightarrow X_k$ on U.

Suppose there is an edge $X_k \to X_i$ on U. $X_k \neq X_j$ because otherwise there is an edge $X_j \to X_i$ in G. Hence X_k is in $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$. But then X_k is not a collider on U, and U does not d-connect X_i to X_j given $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$.

Suppose that the first edge on U is an edge $X_i \to X_k$. It follows that either X_i is an ancestor of X_j, or there is a collider on U. Because G is acyclic, and X_j is an ancestor of X_i, X_i is not an ancestor of X_j. Suppose then that there is a collider X_l on U. Because U d-connects X_i and X_j given $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$, X_l is an ancestor of $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$, and hence of X_i. It follows that G is cyclic, contrary to our assumption that G is a MAG.

Suppose that the first edge on U is $X_i \leftrightarrow X_k$. If there is a collider on U, then X_k is an ancestor of the collider. The collider is an ancestor of $\mathbf{An}(X_i)\backslash\{X_j,X_j\}$ because U d-connects X_i and X_j given $\mathbf{An}(X_i)\backslash\{X_i,X_j\}$. Hence X_k is an ancestor of $\mathbf{An}(X_i)\backslash\{X_i,X_j\}$. It follows that X_k is an ancestor of X_i, contrary to the assumption that G is a MAG. Suppose then that there is no collider on U. Hence X_k is an ancestor of X_j. But X_j is by hypothesis a member of $\mathbf{An}(X_i)$. It follows that X_k is an ancestor of X_i, contrary to the assumption that G is a MAG. $\therefore$

Lemma 6: In a MAG G, if an undirected path U in G d-connects A and B given $\mathbf{An}(A) \cup \mathbf{An}(B)\backslash\{A,B\}$ then U is an inducing path between A and B.

Proof. If there is a path U that d-connects A and B given $\mathbf{An}(A) \cup \mathbf{An}(B)\backslash\{A,B\}$ then every collider on U is an ancestor of a member of $\mathbf{An}(A) \cup \mathbf{An}(B)\backslash\{A,B\}$, and hence an ancestor of A or B. Every vertex on U is an ancestor a collider on U or an ancestor of A or B; and hence every vertex on U except for the endpoints is in $\mathbf{An}(A) \cup \mathbf{An}(B)\backslash\{A,B\}$. If U d-connects A and B given $\mathbf{An}(A) \cup \mathbf{An}(B)\backslash\{A,B\}$, then every vertex that is on U, except for the endpoints, is a collider. Hence U is an inducing path between A and B. $\therefore$

Theorem 5: If G is a MAG, and Σ is a positive definite covariance matrix such that if X_i and X_j are d-separated given Z in G, then $cov(X_i,X_j|\mathbf{Z}) = 0$, then there is a linear parameterization θ of G such that $\Sigma_{G(\theta)} = \Sigma$.

Proof. By Lemma 3, there is a complete MAG G_C such that G is a subgraph of G_C. By Theorem 4, there is a parameterization θ of G_C such that $\Sigma_{G_C(\theta)} = \Sigma$. We will now show that θ assigns zeroes to every edge that is in G_C but not in G. First consider a directed edge $X_i \rightarrow X_j$ that is in G_C but not in G. By the method of construction of G_C, X_i is an ancestor of X_j in G. Because the edge $X_i \rightarrow X_j$ does not exist in G, by Lemma 5, X_j is d-separated from X_i given $\mathbf{An}(X_j)\backslash\{X_i,X_j\}$ in G. Hence $cov_\Sigma(X_i,X_j|\mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$ by hypothesis.

Because G_C is a MAG, ε_j is uncorrelated with the errors of any ancestor of X_j, and hence uncorrelated with any ancestor of X_j. Hence in θ the coefficients of the ancestors of X_j in the equation for X_j are equal to the partial regression coefficients of X_j on its ancestors. But when X_j is regressed on its ancestors, the partial regression coefficient of X_i in the equation for X_j, is equal to zero when $cov_{G_C(\theta)} (X_i,X_j|\mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$. Hence, if there is no edge $X_i \rightarrow X_j$ in G, Xi and Xj are d-separated given $\mathbf{An}(X_j)\backslash\{X_i,X_j\}$ in G, and by hypothesis $cov_\Sigma(X_i,X_j|\mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$. $cov_{G_C(\theta)}(X_i,X_j|\mathbf{An}(X_j)\backslash\{X_i,X_j\}) = cov_\Sigma(X_i,X_j|\mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$. Hence in θ, the partial regression coefficient of X_i in the equation for X_j, is equal to zero. It follows that we can remove from G_C every directed edge that is in G_C but not in G. Call this graph G_{C_0}.

We now show that G_{C_0} is a MAG. G_{C_0}. G_{C_0} has the same ancestor relations as G_C and G. If $A{*}\!\rightarrow B$ in G_{C_0}, then $A{*}\!\rightarrow B$ in G_C. Because G_C is a MAG, B is not an ancestor of A in G_C. Hence B is not an ancestor of A in G_{C_0}.

Suppose that there is an inducing path between A and B in G_{C_0}. Because G_{C_0} is a subgraph of G_C, there is an inducing path between A and B in G_C. If A is an ancestor of B in G_C, then because every vertex on the inducing path is an ancestor of A or B, every vertex on the inducing path is an ancestor of B. But the predecessor of B on the inducing path is a collider on the inducing path, and hence is not an ancestor of B. It follows that A is not an ancestor of B. Similarly, B is not an ancestor of A. Hence the edge between A and

B in G_C is $A \leftrightarrow B$. It follows that the edge $A \leftrightarrow B$. is in G_{C_0}. By Theorem 1, G_{C_0} is a MAG.

Let θ_0 be the same as θ, except that all of the zero coefficients corresponding to directed edges in G_C but not in G have been removed. θ_0 is a parameterization of G_{C_0}, and $\Sigma_{G_{C(\theta_0)}} = \Sigma_{G_{C(\theta)}} = \Sigma$.

Now consider the double-headed arrows in G_C but not in G. Arrange these pairs on the following order O: $\{X_i,X_j\} < \{X_k,X_m\}$ if $\max(i,j) < \max(k,m)$. This ordering has the following property: If $X_i \leftrightarrow X_j$ precedes $X_k \leftrightarrow X_m$ in the order then $X_k \leftrightarrow X_m$ does not occur on any inducing path between X_i and X_j in G_C. Suppose, for a reductio, that this was not the case. Because X_k and X_m occur on an inducing path between X_i and X_j, each of them is an ancestor of X_i or X_j. Suppose without loss of generality that $k > m$. By the ordering of the variable pairs, $k > \max(i,j)$. By the ordering of the individual variables, X_k is not an ancestor of X_i or X_j. This is a contradiction.

Let G_{C_n} be the graph resulting from removing the first n double-headed arrows that are in G_{C_0} but not in G, in the order in which they occur in O, and θ_n be the parameterization of G_{C_n} that results from removing from θ the parameters corresponding to the edges removed from G_{C_0}. The proof that G_{C_n} is a MAG is by induction on the the number of double-headed arrows removed from G_{C_0}. For zero double-headed arrows removed, we have already shown that G_{C_0} is a MAG, and $\Sigma_{G_{C_n(\theta_n)}} = \Sigma_{G_{C(\theta)}} = \Sigma$. Let the induction hypothesis be that G_{C_n} is a MAG and $\Sigma_{G_{C(\theta_0)}} = \Sigma_{G_{C(\theta)}} = \Sigma$. Let the edge $X_i \leftrightarrow X_j$ be edge n+1 in the ordering O. Suppose that in θ_n, $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = c$. We will now show that $G_{C_{n+1}}$ is a MAG.

No double-headed arrow that is in G_{C_n} but not in G appears on an inducing path between X_i and X_j in G_{C_n}, because the only edges that lie on an inducing path between X_i and X_j in G_{C_n} and that are not in G, occur prior to $X_i \leftrightarrow X_j$ in the ordering, and by the induction hypothesis have already been removed from G_{C_n}. Every directed edge that exists in G_{C_n} also exists in G. Hence if there is an inducing path between X_i and X_j in G_{C_n}, there is an inducing path between X_i and X_j in G. But because G is a MAG, and there is no edge between X_i and X_j in G, there is no inducing path between X_i and X_j in G. It follows that there is no inducing path between X_i and X_j in G_{C_n}, other than the edge $X_i \leftrightarrow X_j$. It follows that in $G_{C_{n+1}}$, if there is no edge between X_i and X_j, then there is no inducing path between X_i and X_j.

$G_{C_{n+1}}$ has the same ancestor relations as G_C and G. If $A{*}\!\rightarrow B$ in G_{Cn+1}, then $A{*}\!\rightarrow B$ in G_C. Because G_C is a MAG, B is not an ancestor of A in G_C. Hence B is not an ancestor of A in $G_{C_{n+1}}$. It follows from Theorem 1 that $G_{C_{n+1}}$ is a MAG.

If there is no edge between X_i and X_j in $G_{C_{n+1}}$, by Theorem 1 there is no inducing path between X_i and X_j in $G_{C_{n+1}}$. By Lemma 6, X_i and X_j are d-separated given $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}$. Hence $cov_{G_{Cn+1}(\theta_{n+1})}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$. By hypothesis, $cov_{G_C(\theta)}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = cov_\Sigma(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\})$. Because the edge $X_i \leftrightarrow X_j$ does not occur in G, X_i and X_j are d-separated given $\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}$ in G, and hence $cov_\Sigma(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) = 0$. By Lemma 4,

$$0 = cov_{G_{Cn}(\theta_n)}(Xi,Xj|An(Xi) \cup An(Xj)\backslash\{Xi,Xj\}) =$$

$$cov_{G_{Cn+1}(\theta_{n+1})}(X_i,X_j|\mathbf{An}(X_i) \cup \mathbf{An}(X_j)\backslash\{X_i,X_j\}) + c = 0 + c$$

It follows that $c = 0$, and hence $\Sigma_{G_{Cn+1}(\theta_{n+1})} = \Sigma_{G_{Cn}(\theta_n)} = \Sigma$. ∴

Lemma 7: The n^{th} derivative of a rational function $f_1(X)/f_2(X)$ (where f_1 and f_2 are polynomials) whose denominator is nowhere 0 on its domain is a rational function, whose denominator is a positive integral power of $f_2(X)$.

Proof. Consider the first derivative. The derivative is equal to $(f_1'(X) * f_2(X) - f_1(X) * f_2'(X))/f_2(X)^2$, where f_1' and f_2' are also polynomials. Hence the derivative is rational and has a denominator that is a positive integral function of $f_2(X)$.

Suppose the n^{th} derivative is a rational function, and the denominator is a positive power of $f_2(X)$, i.e. the n^{th} derivative equals $f_3(X)/f_2(X)^m$, where $f_3(X)$ is a polynomial. It follows that the $n{+}1^{st}$ derivative is a rational function $(f_3'(X) * f_2^n(X) - m * f_2(X)^{n-1} * f_1(X) * f_2'(X))/f_2(X)^{m+1}$, where f_1' and f_3' are also polynomial functions. The denominator is a positive power of $f_2(X)$, and hence is nowhere 0 in its domain. ∴

Let the set of values of natural parameters of a full regular exponential family **S** be denoted by **N**, $\mathbf{S_0}$ be a subfamily of **S**, $\mathbf{N_0}$ be the set of values of the natural parameters of $\mathbf{S_0}$, and if **U** is an open neighborhood in **N**, $\mathbf{S^U}$ be the set of distributions in **S** with parameters in **U**.

Theorem 6: The family of distributions represented by a linear MAG M over a set of k variables is a locally parameterized curved exponential family of dimension equal to k(k+1)/2 minus the number of pairs of variables in M that are not adjacent to each other.

Proof. According to Theorem 4.2.1 in Kass and Vos(1997), a subfamily $\mathbf{S_0}$ of an n-dimensional regular exponential family **S** is a locally parameterized curved exponential family if for each η_0 in $\mathbf{N_0}$ there is an open neighborhood **U** in **N** containing η_0 and a diffeomorphism h: $\mathbf{U} \rightarrow \mathbf{R}^{\mathbf{k}} \times \mathbf{R}^{\mathbf{n-k}}$ such that $\mathbf{S_0^U} = \{P_\eta \text{ in } \mathbf{S^U}\text{:}\ h(\eta) = (\beta,\psi) \text{ and } \psi = 0\}$.

According to Theorem 4, the distributions represented by a given MAG M can be parameterized in the following way. For a given covariance matrix Σ among the **X** variables, regress X_k on the set $\mathbf{P_k} := \{X_j \mid X_j \neq X_k \text{ and } X_j \text{ is an ancestor of } X_k\}$. Let

$$\hat{X}_k = \sum_{X_j \in \mathbf{P}_k} \alpha_{kj} X_j$$

be the linear predictor of X_k on $\mathbf{P_k}$, or 0 if $\mathbf{P_k}$ is empty. Now let $\varepsilon_k := X_k - \hat{X}_k$. The α_{kj} and the non-zero covariances among the ε_k parameterize a MAG. Call this set of parameters **M**.

First we will show that there is a diffeomorphism from the set of covariance matrices Σ of the normal distribution (with zero means) over k variables to **M**. By Corollary A.3 in Kass and Vos, it suffices to show that there is a smooth one-to-one function from Σ to **M**, whose inverse is also smooth.

From Theorem 4 it follows that

$$\mathrm{cov}(\varepsilon_p,\varepsilon_q) = \mathrm{cov}(X_p - \hat{X}_p, X_q - \hat{X}_q) =$$
$$\mathrm{cov}(X_p,X_q) - \mathrm{cov}(X_p,\hat{X}_q) - \mathrm{cov}(\hat{X}_p,X_q) + \mathrm{cov}(\hat{X}_p,\hat{X}_q) =$$
$$\mathrm{cov}(X_p,X_q) - \sum_{X_i \in \mathbf{P}_p} \alpha_{pi}\mathrm{cov}(X_p,X_i) - \sum_{X_j \in \mathbf{P}_q} \alpha_{qj}\mathrm{cov}(X_q,X_j) + \sum_{X_i \in \mathbf{P}_p} \sum_{X_j \in \mathbf{P}_q} \alpha_{pi}\alpha_{qj}\,\mathrm{cov}(X_i,X_j)$$

Each of the α_{kj} is a regression coefficient, and hence a rational function of Σ that is everywhere defined on its domain (because every covariance matrix in Σ is positive definite). $\text{cov}(\varepsilon_p,\varepsilon_q)$ is a rational function of the α_{kj} and Σ that is everywhere defined on its domain (because every covariance matrix in Σ is positive definite). Because **M** is a rational function of Σ that is everywhere defined on its domain, by Lemma 7 there is a smooth function from Σ to **M**.

It was also shown in Theorem 4 that each variable X_k could be written as

$$X_k = \sum_{X_j \in \mathbf{Parents}(X_k)} \alpha_{kj} X_j + \varepsilon_k$$

Hence the function mapping Σ to **M** has an inverse and is one-to-one. In addition, it follows that there is a reduced form for the **X** variables, i.e. they are a rational function of the values of the α_{kj} parameters and the ε variables. Hence Σ is a rational function of **M**. It follows that there is a smooth function from **M** to Σ.

Hence there is a diffeomorphism from Σ to **M**.

There is also a diffeomorphism from **N** to Σ (Kass and Vos, p. 101). The composition of two diffeomophisms is a diffeomorphism (Kass and Vos, p. 101), and hence there is a diffeomorphism from **N** to **M**.

Each family of distributions represented by a MAG can be characterized by setting some subset of the parameters of a complete MAG equal to zero. It follows from Theorem 4.2.1 that the distributions represented by a MAG are a curved exponential family.

Since the dimensionality of the full space of k normal variables with zero mean is equal to k(k+1)/2, the dimensionality of a compete MAG is k(k+1)/2. Let M be an incomplete MAG. By Lemma 3, M has a complete extension M', and the dimensionality of M' is k(k+1)/2. Each parameter in M' that is set to zero (one of the α_{kj}, or a covariance between two error terms ε_k and ε_j) corresponds to a pair of variables in M that are not adjacent. The number of parameters in M is equal to k(k+1)/2 minus the number of parameters in M' set to zero, i.e. k(k+1)/2 minus the number of pairs of variables in M' that are not adjacent to each other. $\therefore$

References

Kass, R. and Vos, P. (1997) *Geometrical Foundations of Asymptotic Inference*, Wiley, NY.

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

Spirtes, P., Richardson, T., Meek, C., Scheines, R. and Glymour, C. (1997) "Using D-separation to Calculate Zero Partial Correlations in Linear Models with Correlated Errors", Technical Report CMU-72-Phil.

Spirtes, P., and Richardson, T. (1996). A Polynomial Time Algorithm For Determining DAG Equivalence in the Presence of Latent Variables and Selection Bias, Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models in Proc. Sixth Conference on Uncertainty in AI. Association for Uncertainty in AI, Inc., Mountain View, CA.